Mediated Talk: An Experiment*

Andreas Blume† Ernest K. Lai‡ Wooyoung Lim§

February 11, 2019

Abstract

Theory suggests that mediation has the potential to improve information sharing. This paper experimentally investigates whether and how this potential can be realized. It is the first such study in a cheap-talk environment. We find that mediation encourages players to use separating strategies. Behavior gravitates toward pooling with direct talk and toward separation with mediated talk. This difference in behavior translates into a moderate payoff advantage of mediated over direct talk. There are systematic departures from the equilibrium prediction, characterized by over-communication by senders in the initial rounds of direct talk, stable under-communication by senders under mediated talk, and over-interpretation (attributing too much information to messages) by receivers under both direct and mediated talk.

Keywords: Sender-Receiver Games, Communication, Mediation, Noisy Channels, Laboratory Experiments

***JEL* classification numbers**: C72; C92; D82; D83

*We are grateful to Shih En Lu and Tiemen Woutersen for helpful suggestions. We have benefitted from comments by seminar audiences at Shanghai University of Finance and Economics and Simon Fraser University. This study is supported by a grant from the Hong Kong Research Grant Council (Grant ECS-699613).

†Department of Economics, University of Arizona. *Email:* ablume@email.arizona.edu

‡Department of Economics, Lehigh University. *Email:* kwl409@lehigh.edu

§Department of Economics, The Hong Kong University of Science and Technology. *Email:* wooyoung@ust.hk

1 Introduction

Information transmission is a ubiquitous part of economic activity and inefficiencies affecting it entail potentially significant costs to society. A principal source of inefficiencies is the incentive for strategic manipulation resulting from divergence of interests among communicating parties (Crawford and Sobel [6]). The extent of these inefficiencies will depend on the rules and protocols that govern communication. In this paper we engage in a *communication design* exercise: we investigate whether introducing a (non-strategic) mediator can improve information transmission.

Mediators can help improve information transmission between a sender and a receiver by garbling the sender's messages. A sender who is reluctant to provide information will be more willing to do so knowing that information is degraded by garbling. The receiver prefers garbled information to no information. Under the right conditions both parties gain.[1]

The potential for mediation to improve information transmission has long been recognized. Forges [9] presents an example of an information transmission game in which communication is entirely ineffective with direct communication, while both sender and receiver can gain from communication via an appropriately chosen mediation scheme. The underlying logic is nicely illustrated by Myerson [16] with a story of a sender and receiver communicating via a messenger pigeon. The sender has the choice of either sending the pigeon or not sending the pigeon. If the pigeon is sent, it gets lost with some probability. There are two sender types, one who prefers to be revealed and another who prefers to be concealed. There is an equilibrium in which the type who prefers to be revealed sends the pigeon and the other does not send the pigeon. When the pigeon does not arrive, the receiver does not know whether it was never sent or was sent but got lost. As a result, when the pigeon does not arrive the receiver remains uncertain about whether he is dealing with one type or the other. The type who prefers to be concealed remains concealed and the type who prefers to be identified manages to get identified at least some of the time. The mediation scheme ensures that the type who would prefer to be revealed is sometimes pooled in with the other type, providing the latter type with deniability.

More recently Goltsman, Hörner, Pavlov and Squintani [11] (GHPS) have taken up the question of mediation in the context of the leading example in Crawford and Sobel [6]. They identify an efficiency bound for optimal mediation. Via the revelation principle (Myerson [15]) this is also the bound for any other communication protocol, including, for example, repeated face-to-face communication as considered by Krishna and Morgan [14]. Blume, Board and Kawamura [2] demonstrate that the GHPS efficiency bound can be attainted in a single round of communication through a simple noisy channel: the sender sends a message to the receiver that goes

[1] The sender could accomplish the garbling of messages without a mediator by adopting an appropriate randomization. The problem is that without a mediator that randomization will typically not be incentive compatible. The mediator has a role because he is committed to a garbling rule.

through with some probability as sent; otherwise the message is replaced by a random draw from some fixed distribution. The equilibria that achieve the efficiency bound with this noisy channel exhibit a structure reminiscent of the messenger pigeon example: (sets of) high types are sometimes pooled with (a set of) low types, and otherwise revealed.

We take mediation to the lab. To our knowledge, this is the first paper that experimentally compares direct with mediated communication. To keep the comparison manageable and induce salient incentives, we use a two-type incentive structure and a mediation rule that closely mirrors the one employed in the messenger pigeon example. Our primary interest is in comparing the outcomes from direct talk, where the receiver observes the sender's message as sent, and mediated talk, where the sender's message is filtered through a noisy channel.

In addition, we are interested in how the available language, the way messages are framed, affects mediation outcomes. A natural frame to adopt with a mediator is that of a direct mechanism. In a direct mechanism the sender makes reports about her type to the mediator, using a language consisting of *declaratives*. Based on those reports, the mediator makes action recommendations to the receiver, using a language consisting of *directives*. In contrast, with direct communication the spaces of sent and received messages are identical. This suggests five treatments, direct talk with declaratives, direct talk with directives, and mediated talk, either through a direct mechanism, or using only declaratives, or only directives.

Our theoretical predictions are derived from augmenting a standard equilibrium analysis with a level-k approach that is anchored at the focal meanings of messages. For direct talk, equilibrium alone predicts pooling, which can be achieved with having messages be independent of the sender's type. With mediated talk, it allows for separation, where the two types send distinct messages, although pooling remains an equilibrium. The level-k analysis is consistent with the equilibrium analysis: at all levels above level 0 players use equilibrium strategies. The level-k prediction refines the equilibrium prediction by singling out exactly one message that will be sent with direct talk and predicting separation with mediated talk.

We find that, in line with the theoretical predictions, the modal observed behavior of both senders and receivers converges to pooling with direct talk and to separation with mediated talk. To the extent that there is equilibrium behavior, the equilibria selected and message use are consistent with our level-k analysis. We see separation rather than pooling with mediated talk and pooling on the predicted message with direct talk. Outcomes, defined as joint distributions over states and actions, gravitate toward pooling with direct talk and toward separation with mediated talk. This difference in behavior translates into a moderate payoff advantage of mediated over direct talk.

The choice of language has no impact on terminal behavior. It does not matter whether players communicate using declaratives or directives, or through a direct mechanism that accepts

declaratives and issues directives. This is consistent with standard equilibrium analysis, for which the available language is irrelevant. The observed irrelevance of which specific language is available is also predicted by our level-k analysis.

Some departures of observed from predicted behavior are worth noting: (1) there is a substantial fraction of senders who separate in the first ten rounds of direct talk, contrary to the equilibrium prediction and the level-k prediction for levels 1 and higher; (2) there is a substantial fraction of senders who pool both in the first and last ten rounds of mediated talk, contrary to both the equilibrium and level-k predictions; (3) under mediated talk there is a substantial and stable fraction of receivers whose behavior cannot be explained by *any* belief about sender strategies; and, (4) substantial fractions of state-action pairs in both direct talk and mediated talk are inconsistent with both pooling and separation. As a result of these departures from predicted behavior, players do not approximate the full benefits of mediation. Receivers, in particular, obtain a lower payoff than their guaranteed expected payoff in both direct and mediated talk. In direct talk they do not even achieve a payoff they could guarantee with certainty.

In summary, the equilibrium analysis accompanied by a level-k analysis anchored at focal message meanings captures that mediation encourages separation, but both miss some other salient features of the data. The level-k analysis, in particular, proves useful for equilibrium selection, accounting for message use, explaining initial separation in direct talk, rationalizing off-equilibrium-path responses under direct talk, but it does not account for departures from equilibrium behavior under mediated talk.

There is a rich literature on sender-receiver game experiments with direct talk. We survey that body of work in Blume, Lai, and Lim [3]. The experimental papers most closely related to ours are Nguyen [17], Blume, Lai, and Lim [4], and Fréchette, Lizzeri, Perego [10]. Nguyen and Fréchette et al. experimentally investigate Bayesian persuasion (Kamenica and Gentzkow [13]). With Bayesian persuasion, like in the present paper, the receiver observes a garbled signal of the state of the world; unlike here, there is no private information and the sender can fully commit to a signaling rule that maps states of the world into signals. Blume et al. [4] model and implement Warner's [19] randomized response method in the lab. Under randomized response, garbling is entirely under the control of the sender, and for that to be incentive compatible it is necessary that the sender has a preference for compliance with the procedure, which may be in the form of deriving utility from truth-telling. While messages in the present paper are cheap talk, communication under randomized response amounts to costly signaling. Ours is the first paper that looks at the effects of garbling cheap-talk messages.

In the next section we introduce the communication protocols and the incentive structure. In Section 3 we discuss the theoretical predictions. Section 4 describes our experimental treatments and procedures. In Section 5 we report our findings and in Section 6 we discuss our findings and

possible extensions.

2 The Communication Environments

We investigate the impact of mediation on communication between a privately informed sender and an uninformed receiver whose action determines the payoff of both players. In mediated talk the sender sends a message to a non-strategic mediator who in turn sends a message to the receiver. The receiver takes an action after having observed (only) the mediator's message. We contrast mediated talk with direct talk, where the sender sends a message to the receiver without the intervention of a mediator.

A mediation scheme that has played a central role in theory and is natural for our environment is a *direct mechanism*. In a direct mechanism, the sender's message to the mediator is framed as a declaration of her private information, her type, and the mediator's message to the receiver is framed as a directive of which action to take. Therefore, with a direct mechanism we have two message spaces to consider, one composed of *declaratives* and one composed of *directives*.

In general, with both mediated and direct talk we have a choice of message spaces. To allow for the possibility that the framing of messages matters, we consider two direct-talk treatments, one with declaratives and one with directives, and three mediated-talk treatments, one in which the mediator both receives and sends declaratives, one in which the mediator both receives and sends directives, and the direct mechanism, in which the mediator receives declaratives and sends directives. The size of the message spaces is the same across all treatments, and is the minimal size required to induce all possible equilibrium outcomes. Thus in total we have five treatments, two for direct talk and three for mediated talk.

In all five games considered the sender privately observes the state of the world and then sends a message. After observing a possibly garbled (through the mediator) version of the sender's message the receiver takes an action. Payoffs of both players depend on the state of the world and the action taken by the receiver. Messages have no direct effect on payoffs, and are thus cheap talk.

	L	C	R
s	$110, 120$	$10, 0$	$60, 100$
t	$80, 0$	$10, 120$	$130, 90$

Table 1: Payoffs

The payoff structure is the same across all games considered. There are two possible states

of the world, s and t, which are commonly known to be equally likely. The receiver has three actions L, C and R. Payoffs are summarized in Table 1. In each cell of the payoff table the first entry indicates the sender's payoff and the second entry the receiver's payoff. This incentive structure captures two central features of the one made prominent by Crawford and Sobel [6]: there are possible efficiency gains from communication and there are incentives for the sender to misrepresent her type. Without communication the receiver's unique optimal action is R. Both sender and receiver could gain if the sender could credibly reveal when her type is s. There is, however, an incentive problem because if type s credibly reveals, then type t prefers to mimic s rather than be revealed herself.

With direct talk, sent and received messages coincide. We consider two direct-talk games, the direct-declaratives game and the direct-directives game.

Table 2: Direct-Talk Games

Game	Direct declaratives	Direct directives
Transmission Rule (input $\overset{\text{prob.}}{\longrightarrow}$ output)	$s \xrightarrow{1} s$ $t \xrightarrow{1} t$	$L \xrightarrow{1} L$ $R \xrightarrow{1} R$

In the direct-declaratives game the message space is $\{s, t\}$, and thus coincides with the type space. Furthermore, if the sender sends message s, the receiver observes message s, and similarly for message t. Sometimes it will be convenient to differentiate messages from types by enclosing them in quotes, e.g. to write "s" and "t" for messages.

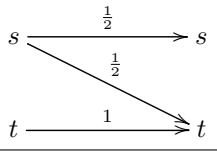
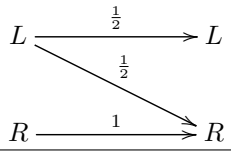
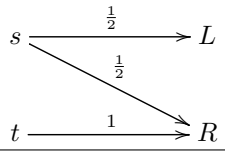
In the direct-directives game the message space is $\{L, R\}$, and thus corresponds to a subset of the receiver's action space.[2] If the sender sends message L, the receiver observes message L, and similarly for message R. Again, for the sake of differentiating messages from actions, we will sometime write "L" and "R" for messages. It will also be convenient sometimes to lump messages "s" and "L" together and to write "s/L". Similarly we will sometimes use "t/R" when we want to talk about messages "t" and "R" at the same time.

With mediated talk, messages pass through a noisy channel and therefore the message observed by the receiver need not coincide with the message sent by the sender. We consider three mediated-talk games, the mediated-declaratives game, the mediated-directives game, and the mediated-direct-mechanism game.

In the mediated-declaratives game the message space is $\{s, t\}$, and thus coincides with the type space. The message observed by the receiver, however, need not be the one sent by the

[2]There is no loss from excluding a possible message C, since it has no effect on the set of equilibrium outcomes. Also, there is no plausible use of such a message that would be in line with the focal interpretation of messages.

Table 3: Mediated-Talk Games

Game	Mediated declaratives	Mediated directives	Mediated direct mechanism
Mediation Rule (input $\overset{\text{prob.}}{\longrightarrow}$ output)			

sender. Instead, when receiving message s from the sender, the mediator randomizes uniformly over messages s and t, when sending a message to the receiver. When receiving message t from the sender, the mediator passes it on to the receiver without change. This mechanism has the structure of an optimal mechanism in our environment; the receiver randomizes after one message and faithfully transmits the other.[3]

In the mediated-directives game the message space is $\{L, R\}$, a subset of the receiver's action space. The mediator in the mediated-directives game uses a mediation scheme that closely mirrors that used in the mediated-declaratives game: when receiving message L from the sender, the mediator randomizes uniformly over messages L and R, when sending a message to the receiver. When receiving message R from the sender, the mediator passes it on to the receiver without change. As before, this mechanism has the structure of an optimal mechanism.

In the mediated-direct-mechanism game the sender sends messages from the message space $\{s, t\}$ to the mediator, and the mediator transmits messages from the message space $\{L, R\}$ to the receiver. Up to relabeling of messages, the mediation scheme is the same as before and thus has the structure of an optimal mechanism: when receiving message s from the sender, the mediator randomizes uniformly over messages L and R; when receiving message t from the sender, the mediator sends message R to the receiver.

3 Theoretical Predictions

In a cheap-talk game the outcome associated with a (Bayesian Nash) equilibrium is the joint distribution over types and actions that is induced by that equilibrium. Our predictions are based on a full characterization of the set of equilibrium outcomes for each game, supplemented by a level-k analysis.

Cheap talk generally admits multiple equilibria supporting the same outcome, due to exchangeability of messages. With mediation, given the specific structure of each mediation scheme,

[3] An optimal mechanism differs only in the probabilities with which the mediator randomizes following message s from the sender. In an optimal mechanism the receiver would be indifferent between actions C and R following message t from the mediator. We chose to avoid this indifference in order to help making incentives salient in the experiment.

this effect is somewhat muted, but even there does not fully disappear. The level-k analysis has a twofold role in our analysis. It admits boundedly rational behavior and it permits predictions to be anchored in the focal interpretations of messages, the "language" we employ in each game. Anchoring predictions in the available language turns out to lead to a unique prediction for each game.

We follow Crawford's [7] proposal for how to handle communication games with focal message meanings in a level-k framework (see also Cai and Wang's [5] application to CS-type sender-receiver games and Wang, Spezio, and Camerer's [18] support for the level-k model in sender-receiver games via eyetracking). The key is that in communication games with focal message meanings those meanings are a natural anchor for players' reasoning. Specifically, we model level-0 (L_0) senders as being *forthright* and L_0 receivers as being *credulous.* A forthright sender is truthful when messages are framed as declaratives and recommends her preferred action when messages are framed as directives. A credulous receiver best-responds to a forthright sender. For higher levels of sophistication, we have $L_{k\geq 1}$ senders (levels k that are at least $k = 1$) best-respond to L_{k-1} receivers and $L_{k\geq 1}$ receivers best-respond to L_k senders. We further postulate that senders who are indifferent between the two messages and receivers who encounter an unexpected message behave as they would at the next lower level.[4]

As we will see, in each game the set of equilibrium outcomes is very small. The level-k analysis refines this prediction further and selects a single equilibrium for each case.

3.1 Direct Talk

With direct talk *separation*, where different types of the sender send distinct messages, is not an equilibrium outcome. Type t of the sender would receive her lowest possible payoff, 10, with separation. She would be better off mimicking type s for a payoff of 80. This breaks any candidate for a separating equilibrium.

Like in any sender-receiver game, with direct talk *pooling* is an equilibrium outcome. Under pooling regardless of the state of the world the receiver takes the action that is optimal given his prior beliefs, here action R. The joint distribution over types and actions that corresponds to the pooling outcome is shown in Table 4. It reflects the fact that types are equally likely and that the receiver takes action R irrespective of the type. We show in the appendix that ***under***

[4]Our setup introduces two aspects that are potentially relevant for a level-k analysis and not present in Crawford [7]: messages are sometimes framed as directives or garbled by a mediator. As a result, with directives even an unsophisticated sender needs to pay attention to payoffs if we want to connect her message use to focal message meanings. In addition, with directives rather than expressing receiver credulity by having L_0 receivers best respond to a forthright sender strategy, we could have L_0 receivers take directives at face value. Finally, a decision has to be made on how credulous receivers, if they best respond to forthright senders, deal with mediation. Our modeling choices are motivated by trying to stay close to Crawford's original formulation, maximizing the predictive power of our level-k analysis, and the desire to provide a good fit of our data.

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Table 4: Pooling

direct talk pooling is the unique equilibrium outcome.

The equilibrium analysis remains silent about how messages will be used in equilibrium under direct talk. There are equilibria in which the sender sends messages s regardless of type, equilibria in which she sends messages t regardless of type, as well as a continuum of equilibria with different forms of randomization. The level-k analysis can be more precise because it makes use of focal message meanings.

Table 5: Level-k Prediction for the Direct-Declaratives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	C
$L_{k\geq 1}$	"s"	"s"	R	C

Table 5 summarizes the level-k prediction for the direct-declaratives game. With the message space $\{s, t\}$, forthright L_0 senders are truthful. Credulous L_0 receivers take senders to be truthful and respond accordingly. Senders at level L_1, who best respond to L_0 receivers all strictly prefer to send message s, regardless of their true type. Therefore, since message s is uninformative, L_1 receivers respond with action R. This pattern of behavior is stable and iterates through all higher levels: at all but the lowest level of sophistication the prediction is that senders send message s, receivers respond with the pooling action R to message s, and with action C to message t. Hence, the level-k analysis selects a unique pooling equilibrium: at all but the lowest level players employ the same pooling-equilibrium strategy. ***For observables in the direct-declaratives game the level-k prediction is that the sender always sends message s and that the receiver responds to message s with action R.***

The level-k prediction for the direct-directives game is summarized in Table 6. With message space $\{L, R\}$ (forthright) L_0 senders ask for their preferred action. L_0 receivers trust that senders will be forthright and respond accordingly. At all levels other than the lowest level senders pool on message L and receivers respond to message L with action R and to message R with action C. ***For observables in the direct-directives game the level-k prediction is that the sender always sends message L and that the receiver responds to message L with action R.***

Table 6: Level-k Prediction for the Direct-Directives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"L"	"R"	L	C
$L_{k\geq 1}$	"L"	"L"	R	C

3.2 Mediated Talk

With mediated talk *pooling* remains an equilibrium outcome and there are multiple equilibria supporting that outcome. Unlike with direct talk, with mediated talk *separation*, where the sender uses a separating strategy and the receiver best responds to the mediator's induced messages, is an equilibrium outcome. To see this, consider the mediated-direct-mechanism game (the argument applies to the other mediated-talk games with the appropriate relabeling of messages). When the sender reports truthfully, i.e. sends s to the mediator when her type is s and likewise sends t when her type is t, the receiver assigns posterior probability 1 to type s after receiving message L from the mediator and assigns posterior probability $\frac{1}{3}$ to type s after receiving message R from the mediator. This makes it a (unique) best reply for the receiver to take action L after message L from the mediator and to take action R after message R from the mediator. Given that strategy of the receiver, it is uniquely optimal for the sender to report truthfully.[5] The equilibrium outcome in which there is separation is shown in Table 7.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Table 7: Separation

We demonstrate in the appendix that ***with mediation separation and pooling are the only two equilibrium outcomes.*** Furthermore, ***there is a unique equilibrium supporting separation***, which implies that ***for the case of separation, the equilibrium analysis pins down message use.*** For the case of pooling, the equilibrium analysis does not pin down message use.

Separation under mediation also has attractive stability properties: It is supported by the

[5]The structure of this equilibrium is reminiscent of that of optimal equilibria in Blume, Board, and Kawamura [2]. They analyze communication through a noisy channel in the uniform-quadratic version of the Crawford-Sobel [6] model. Equilibria that attain the efficiency bound established by Goltsman, Hörner, Pavlov and Squintani [11] have intervals of high types be revealed some fraction of the time and otherwise pooled with the lowest interval of types. In both cases, this provides a cover for low types. Here the fact that s is sometimes pooled with t protects type t from receiving her least favorite actions C.

unique persistent equilibrium (Kalai and Samet [12]) in the game (which is also the unique minimal curb equilibrium, Basu and Weibull [1]). Taking into account these stability considerations ***our equilibrium analysis predicts* separation *with mediation.*** Separation Pareto dominates pooling both *ex ante* and *ex post*: type s is strictly better off and type t is no worse off under separation.

Additional support for the separation prediction with mediation comes from the level-k analysis. Tables 8-10 report the level-k analysis for the three mediated-talk games. In all three cases forthrightness of the sender at the lowest level translates into separation, which is preserved through all levels.

Table 8: Level-k Prediction for the Mediated-Declaratives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	R
$L_{k\geq 1}$	"s"	"t"	L	R

Table 9: Level-k Prediction for the Mediated-Directives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"L"	"R"	L	R
$L_{k\geq 1}$	"L"	"R"	L	R

Table 10: Level-k Prediction for the Mediated-Direct-Mechanism Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"s"	"t"	L	R
$L_{k\geq 1}$	"s"	"t"	L	R

The level-k analysis predicts that in the mediated-declaratives game senders are truthful and receivers respond to message s with action L and to message t with action R, in the mediated-declaratives game senders sincerely ask for their favorite action and receivers respond to message L with action L and to message R with action R, and in the mediated-direct-mechanism game senders are truthful and receivers respond to message L with action L and to message R with action R.

4 Experimental Treatments and Procedures

Each of the five games analyzed above corresponded to an experimental treatment, as summarized in Table 11.

Table 11: Experimental Treatments

	Direct Talk	Mediated Talk
Declaratives	*Direct-Declaratives*	*Mediated-Declaratives*
Directives	*Direct-Directives*	*Mediated-Directives*
Direct Mechanism	N/A	*Mediated-Direct-Mechanism*

Our experiment was conducted in English using z-Tree (Fischbacher [8]) at The Hong Kong University of Science and Technology. A total of 456 subjects participated in the five treatments. Subjects had no prior experience with the experiment and were recruited from the undergraduate population of the university.

Five sessions were conducted for each treatment. On average, 18 subjects participated in a session, with half of them randomly assigned to the role of a Sender and the other half to the role of a Receiver.[6] Roles remained fixed throughout a session. Senders and receivers in a session were *randomly matched* to play 60 rounds of the game.

In each session, upon arrival, subjects were instructed to sit at separate computer terminals. Each received a copy of the experimental instructions. The instructions were read aloud using slide illustrations as an aid. A comprehension quiz and a practice round followed.

Subjects were told that there would be two equally likely situations, situation s and situation t, and their rewards in the two situations differed according to Table 1, which was shown on their decision screens. At the beginning of each round, the computer randomly selected a situation. The sender privately learned the selected situation. In the mediated-direct-mechanism and the two declaratives treatments, the sender chose one of two messages, "s" or "t," to send to the receiver. In the two directives treatments, the choices were messages "L" and "R."

In the direct-talk treatments, the sender's chosen message was always transmitted to the paired receiver as sent. In the mediated-declaratives or directives treatment, message "t/R" chosen by the sender was always transmitted to the receiver as sent, while for message "s/L"

[6] Of the 25 sessions, there were 13 with 20 subjects, 6 with 18 subjects, 3 with 16 subjects, 2 with 14 subjects, and 1 with 12 subjects.

there were equal probabilities that "s/L" or "t/R" was transmitted. In the mediated-direct-mechanism treatment, "t" was always transmitted as "L," and "s" was transmitted as "L" or "R" with equal probabilities. Subjects were informed of these mediation rules.

After receiving a message, the receiver chose one of three actions, L, C, or R. Rewards for the round were then determined based on the randomly selected situation and the receiver's action. Feedback on the situation, the sender's message, the receiver's action, and the subject's reward was provided at the end of each round.

We randomly selected two rounds for payments. The average reward a subject earned in the two selected rounds was converted into Hong Kong Dollars at a fixed and known exchange rate of HK$1 per reward point. A show-up fee of HK$30 was also provided. Subjects on average earned HK$119.3 (≈ US$15.3) by participating in a session that lasted 1.6 hours.

5 Findings

For each of our games, level-k behavior for $k \geq 1$ coincides with behavior in an equilibrium. When contrasting observed behavior with predictions, "prediction" refers to this equilibrium.

5.1 Mediated vs direct talk

In this section we compare the data from mediated talk, aggregated over all sessions of the three mediated-talk treatment, with the data from direct talk, aggregated over all sessions of the two direct-talk treatments. We focus on terminal behavior (last ten rounds) and initial behavior (first ten rounds).

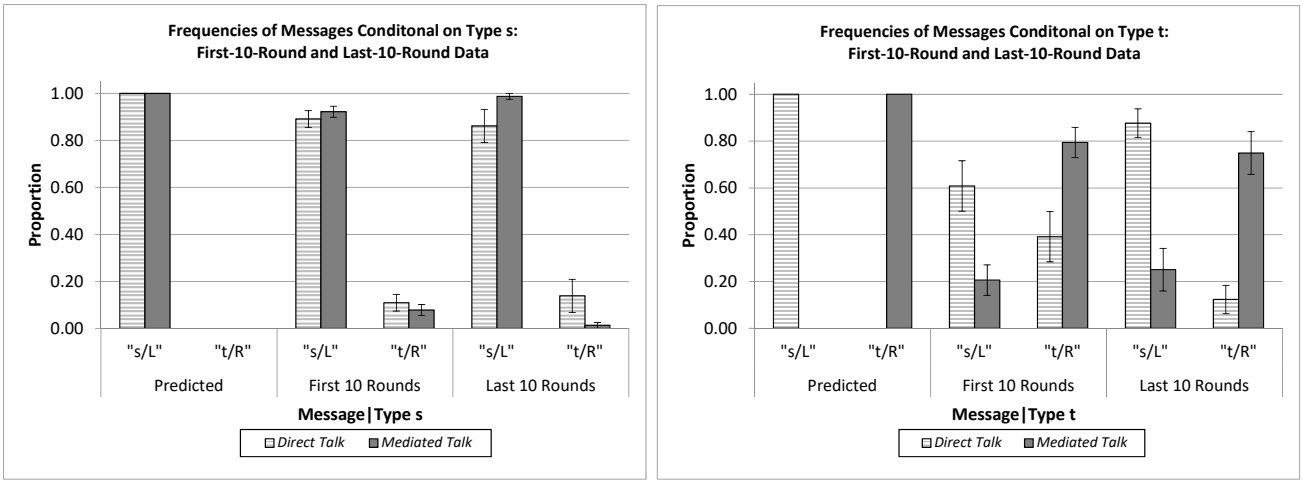
5.1.1 Mediated vs direct talk - sender behavior

Figure 1 summarizes sender behavior in the mediated-talk and direct-talk treatments over the first and last ten rounds, and compares observed with predicted behavior (error bars correspond to 95% confidence intervals). It indicates for each type, s and t, of the sender the frequencies with which that type sends message "s" and message "t", in case the message space is {"s", "t"}, or message "L" and message "R", in case the message space is {"L", "R"}. Since for all our treatments the theoretical predictions do not distinguish messages "s" and "L", we lump them together where appropriate, and refer to messages "s/L"; similarly, we use "t/R" to jointly refer to messages "t" and "R". The principal difference in predictions between the direct and mediated-talk treatments is that with direct talk type t pools with type s on message "s/L" whereas with mediated talk type t separates by sending message "t/R". The main takeaway from Figure 1 is

that, as predicted, there is more sender separation with mediated talk than with direct talk, and modal message use tends to conform with the theoretical prediction.

In the last 10 rounds of the direct-talk treatments, type-t senders send message "s/L" 88% of the time, significantly more often than message "t/R" ($p = 0.003$, Wilcoxon signed-rank test). By contrast, in the last 10 rounds of the mediated-talk treatments, type-t senders send message "t/R" 75% of the time, significantly more often than message "s/L" ($p < 0.001$, Wilcoxon signed-rank test). For comparison across the two sets of treatments, type t-senders in the mediated-talk treatments separate by sending "t/R" significantly more often than do type t-senders in the direct-talk treatments ($p < 0.001$, Mann-Whitney test).

Quantitatively, the differentiation between the two sets of treatments is overall less pronounced in the first 10 rounds, but the differences remain statistically significant. In the first 10 rounds of the direct-talk treatments, type-t senders send message "s/L" 61% of the time, significantly more often than message "t/R" ($p = 0.02$, Wilcoxon signed-rank test). In the last 10 rounds of the mediated-talk treatments, type-t senders send message "t/R" 79% of the time, significantly more often than message "s/L" ($p < 0.001$, Wilcoxon signed-rank test). It also remains the case that type t-senders in the mediated-talk treatments separate by sending "t/R" significantly more often than do type t-senders in the direct-talk treatments ($p < 0.001$, Mann-Whitney test).



(a) Type s (b) Type t

Figure 1: Senders' Behavior in Direct Talk vs. Mediated Talk:
First-10-Round and Last-10-Round Data

In the last ten rounds with direct talk the behavior of type t-senders is indistinguishable from that of type s-senders: the frequency of type-s senders sending "s/L" is 86%, while that of type-t sending "s/L" is 88%. This tendency to pool in the terminal periods evolves over time from type-t senders sometimes separating during the initial 10 rounds: in the first 10 rounds

of the direct-talk treatments, type-t senders send message "t/R" 39% of the time. Thus, while initially there is over-communication with direct talk, it disappears over time. Terminal play in direct talk converges toward types pooling, and, as predicted by the level-k analysis, pooling on the common message "s/L".

With mediated talk, the predominant sender behavior is separation, where, consistent with the theoretical prediction for how messages are used, type s senders send message "s/L" and type t sender send "t/R". Separation starts early and does not vary much over time. Senders of type s send message "s/L" 92% of the time in the first 10 rounds and 99% of the time in the last 10 rounds. Senders of type t send message "t/R" 79% of the time in the first 10 rounds and 75% of the time in the last 10 rounds.

The principal departure from predicted sender behavior is that a non-negligible fraction of t-types fails to separate under mediated talk: senders of type t send message "s/L" 21% of the time in the first 10 rounds and 25% of the time in the last 10 rounds. Despite this, there is a substantial difference between direct and mediated talk. Under direct talk, t-types separate (by sending message "t/R") only 12% of the time in the last 10 rounds. In contrast, under mediated talk t-types separate 75% of the time in the last 10 rounds. The key message regarding sender behavior is therefore that while only a minority of t-types separate under direct talk, most t-types separate under mediated talk.

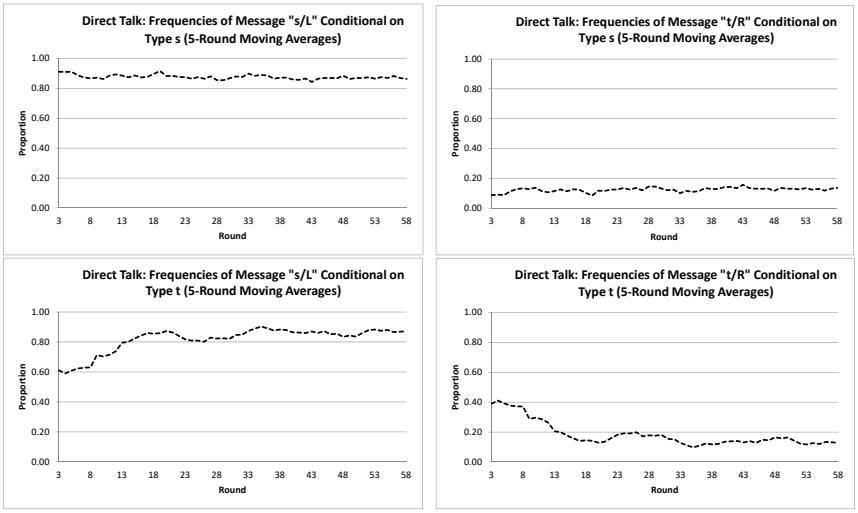
Figure 2 documents the sender dynamics. The figure presents the 5-round moving averages of the frequencies of messages "s/L" and "t/R" conditional on type s and type t in the direct-talk and the mediated-talk treatments.[7]

The figure shows that under direct talk, type s senders send message "s/L" with a high and stable frequency throughout. Type t senders likewise send message "s/L" with higher frequency than message "t/R" from the beginning. While initially there is a substantial fraction of type t senders sending message "t/R", that fraction gradually diminishes until it stabilizes to a level below 20%. In the final 30 rounds both types s and t consistently send message "s/L" with a frequency above 80%. This corresponds to a pooling strategy for the sender: the receiver's best reply to this behavior is to take action R after both messages. Furthermore in the last 30 rounds 80% of sender behavior is consistent with the predicted sender strategy of sending message "s/L" regardless of type.

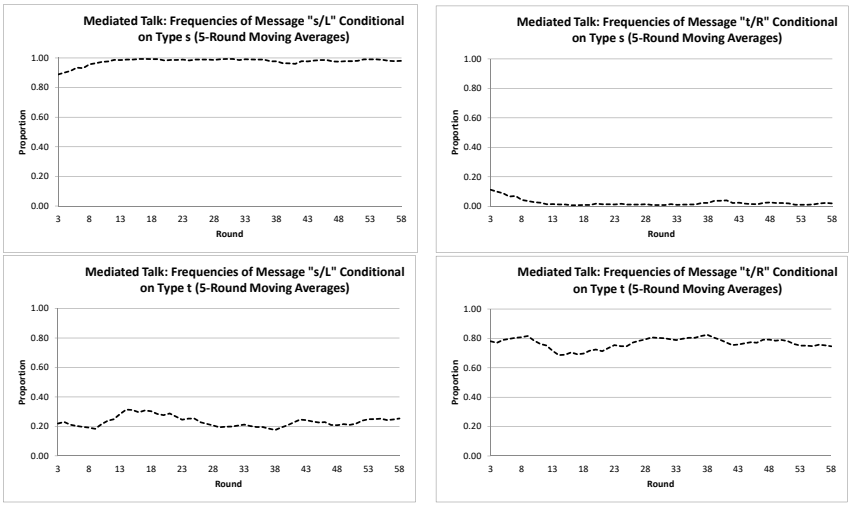
Under mediated talk, after a brief adjustment, nearly 100% of types s consistently send message "s/L", and just above 70% of types t send message "t/R" throughout. While a non-negligible fraction of type t senders pools with type s by sending message "s/L", above 75% send the separating message "t/R" during the last 30 rounds. Overall, behavior under direct

[7]The moving average for round n is calculated by averaging the frequencies in rounds $n-2$, $n-1$, n, $n+1$, and $n+2$. The data points accordingly start at Round 3 and end at Round 58.

talk gravitates toward pooling, whereas behavior under mediated talk is more consistent with separation.



(a) Direct-Talk Treatments



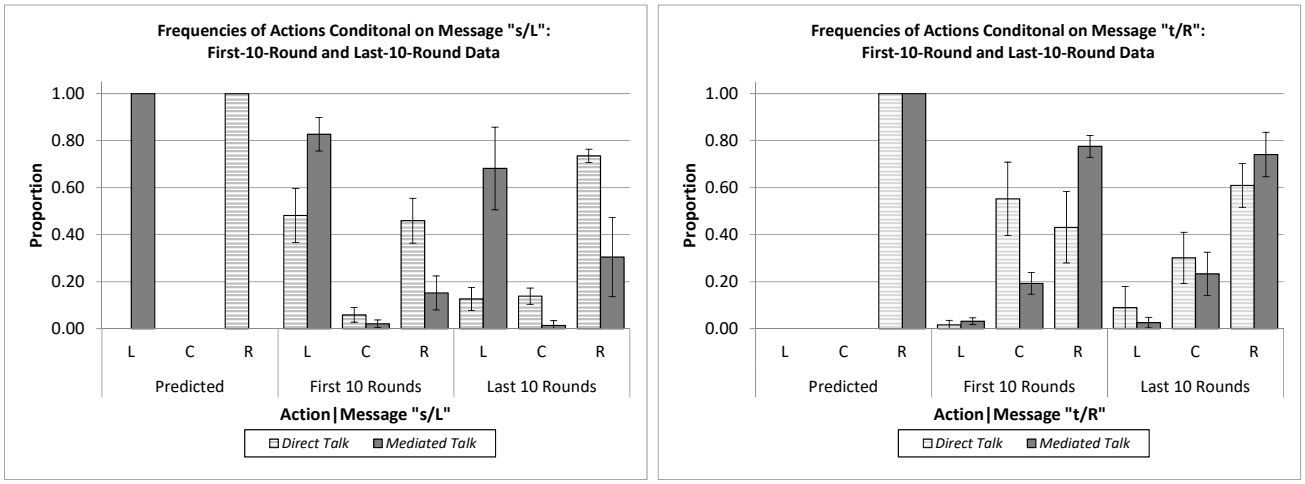
(b) Mediated-Talk Treatments

Figure 2: Trends of Frequencies of Messages Conditional on Types (5-Round Moving Averages)

5.1.2 Mediated vs direct talk - receiver behavior

Figure 3 summarizes initial and terminal receiver behavior in the mediated-talk and direct-talk treatments. Like for senders, modal behavior in terminal rounds gravitates toward pooling with direct talk and toward separation with mediated talk.

In the last 10 rounds of the direct-talk treatments the frequency of the pooling action R is 72%, up from 46% during the first 10 rounds. This increase is seen after both messages. Conditional on message "s/L", the frequency of R is 74% in the last 10 rounds, up from 46%



(a) Message "s/L" (b) Message "t/R"

Figure 3: Receivers' Behavior in Direct Talk vs. Mediated Talk:
First-10-Round and Last-10-Round Data

in the first 10 rounds; conditional on message "t/R", the frequency of R is 61% in the last 10 rounds, up from 43% in the first 10 rounds. During the first 10 rounds, the frequency of action L after message "s/L" is 48%, and after message "t/R" the most frequent action is C, with a frequency of 55%. Thus initial receiver behavior under direct talk is more consistent with separation. This departure from the theoretical prediction is attenuated over time, with pooling eventually becoming the modal behavior under direct talk.

In the last 10 rounds of the mediated-talk treatments 72% of receiver behavior is most consistent with separation. This behavior is established early, with 79% of receiver responses being consistent with separation during the first 10 rounds. There are two kinds of noteworthy departures from theory under mediated talk in the last 10 rounds: the frequency of action R conditional on message "s/L" is 30%, while the frequency of action C conditional on message "t/R" is 23%. Nevertheless, in line with theory, receiver behavior consistent with separation is significantly more pronounced with mediated talk than with direct talk: while under direct talk receivers respond to message "s/L" with the separating action L only 13% of the time in the last 10 rounds, under mediated talk this frequency rises significantly to 68% ($p < 0.001$, Mann-Whitney test). This frequency is observed at different levels in the first 10 rounds, but there is a similar significant difference between direct talk and mediated talk; the receivers respond to message "s/L" with action L 48% of the time in the first 10 rounds under direct talk, while this frequency rises to 83% under mediated talk ($p < 0.001$, Mann-Whitney test).

Figure 4 documents the receiver dynamics. The figure presents the 5-round moving averages of the frequencies of actions L, C and R conditional on message "s/L" and message "t/R" in the direct-talk and the mediated-talk treatments.

The figure shows that under direct talk after both messages the frequency of action R gradually rises, starting at around 40%. Conditional on message "s/L", eventually about 70% of actions taken are action R. Conditional on message "t/R", the trend is slightly less pronounced with eventually just above 60% of actions taken being action R. At least initially the bulk of the remaining responses can be attributed to credulous receiver behavior, where the receiver best responds to assuming that the sender either honestly declares her type or is honest about her most preferred action.

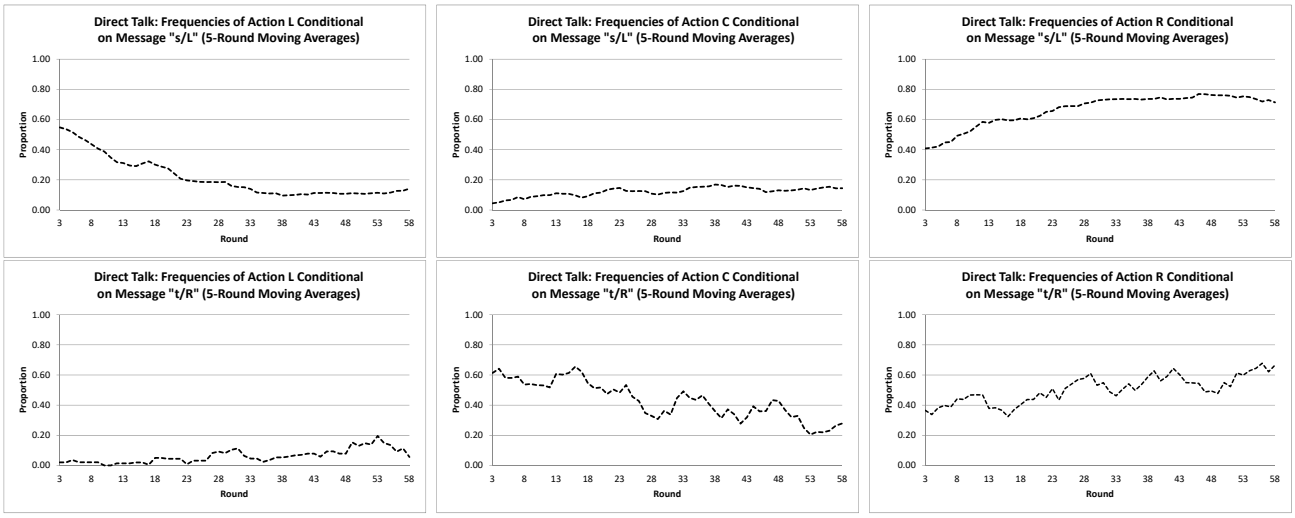
Under mediated talk, there is a stable tendency toward separation throughout. After message "s/L", initially roughly 80% of the actions taken are action L; that frequency drops slightly over time, stabilizing at around 70% during the last 30 rounds. After message "t/R", the frequency of action R is consistently above 70% over the entire 60 rounds. Departures from separation are likewise stable and in two directions. Following message "s/L" in excess of 20% of actions taken are action R during the last 30 rounds and following message "t/R" about 20% of the actions taken are action C over the entire 60 rounds.

The latter departure from the theoretical prediction is of particular interest because it cannot be justified by *any* receiver belief about the sender's strategy. There does not exist a sender strategy that would make it optimal for the receiver to take action C following message "t/R". Any receiver strategy that prescribes action C in response to message "t/R" is strictly dominated, therefore fails to be rationalizable, and cannot be accounted for by a level-k analysis based on receivers fully understanding and internalizing the rules of mediated communication.

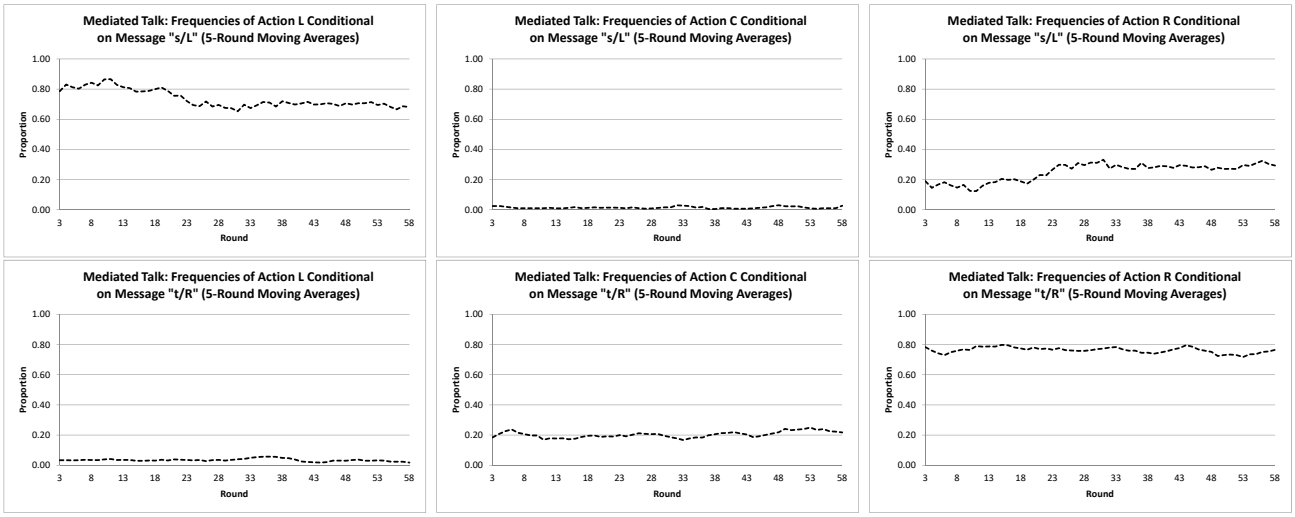
With direct talk, responding with action C to message "t/R" would be rationalizable and is in fact the predicted behavior according to our level-k analysis of direct talk. Therefore observing this behavior under mediated talk would be consistent with a fraction of receivers not understanding the implications of mediation and instead analyzing the game as if it were direct talk.[8] The receiver behavior we observe under mediation is consistent with about 20% of receivers analyzing the game as if there is no mediation. The same rationale, that there is a fraction of about 20% of players who approach mediated talk as if it is direct talk would account for the systematic component of the departure of observed sender behavior from predicted sender behavior under mediated talk that is illustrated in Figure 2.

Overall, we find that, as for senders, receiver behavior under direct talk gravitates toward pooling, whereas receiver behavior under mediated talk is more consistent with separation. We also again find systematic departures from this central tendency, which might be explained by a fraction of players confusing mediated with direct talk.

[8]We show in the appendix that this perspective is consistent with the remaining receivers, who understand that there is mediation, analyzing the game as before, even when they take into account that there is a small fraction (less than 2/9) of players who do not understand that there is mediation.



(a) Direct-Talk Treatments



(b) Mediated-Talk Treatments

Figure 4: Trends of Frequencies of Actions Conditional on Messages (5-Round Moving Averages)

5.1.3 Mediated vs direct talk - outcomes

Table 12 reports outcomes – joint distributions over types and actions – in the mediated-talk and direct-talk treatments over the first and last ten rounds, and relates them to predicted behavior. Predicted outcomes and observations consistent with those outcomes are indicated in bold. Theory predicts that with mediation conditional on type s being realized that type is identified half of the time with mediated talk and never identified with direct talk. Identification of type s is possible under mediated talk because there the sender uses a separating strategy. The principal message from our data is that there is indeed a higher incidence of separation (type-s identification) under mediated talk than under direct talk.

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	22%	6%	**24%**
t	14%	12%	**22%**

First 10 Rounds

	L	C	R
s	6%	8%	**38%**
t	6%	8%	**34%**

Last 10 Rounds

(a) Direct Talk

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	**18%**	6%	**24%**
t	5%	9%	**38%**

First 10 Rounds

	L	C	R
s	**18%**	7%	**27%**
t	4%	9%	**35%**

Last 10 Rounds

(b) Mediated Talk

Table 12: Communication Outcomes: Joint Frequencies over Types and Actions in the First and Last 10 Rounds

In the last ten rounds with direct talk the sum of the frequencies of the pairs (s, R) and (t, R), which are the ones consistent with the pooling outcome, is 72%. The initial frequency of observations consistent with pooling is 46%, with departures from pooling consistent with some separation and some success of type t mimicking type s. In the first 10 rounds under direct talk, the frequency of sender types being identified is 34%, which is the sum of the frequency of the pair (s, L) at 22% and that of the pair (t, C) at 12%. The frequency of the pair (t, L), which indicates that the t-type successfully mimics the s-type, is 14%. By the end, departures from pooling are unsystematic, and pooling emerges as the most frequently observed outcome.

Under mediated talk in both the last and first ten rounds the sum of the frequencies of the pairs (s, L), (s, R), and (t, R), which are the ones consistent with separation, is 80%. There is little structure in the departures from the theoretical outcome prediction under mediated talk both during the last ten rounds and first ten rounds; the most frequent departure is toward (t, C), which is observed with a frequency of 9%.

Consistent with the theory favoring separation under mediated talk, the fraction of observations of (s, L) pairs during the last ten rounds is 18% with mediated talk, whereas it is only 6% with direct talk. The frequency of action L *conditional* on type s is 13% in the last 10 rounds under direct talk, significantly lower than the 35% under mediated talk ($p < 0.001$, Mann-Whitney test). By contrast, there is no significant difference between the conditional frequencies under direct and mediated talk in the first 10 rounds, which are, respectively, 42% and 37% (two-sided $p = 0.4052$, Mann-Whitney test). The central message regarding the outcome comparison is that while initial attempts at communication gradually break down under the weight of incentives

with direct talk, they persist with mediated talk.

Figure 5 documents the outcome dynamics. The figure presents the 5-round moving averages of the frequencies of actions L, C and R conditional on type s and type t in the direct-talk and the mediated-talk treatments.

The figure shows that under direct talk for both types the conditional frequency of action R gradually rises from a starting point of around 40%. Conditional on either type, eventually about 70% of actions taken are action R. Initial departures from the pooling outcomes are in the direction of separation. Over time, departures from pooling become largely unsystematic.

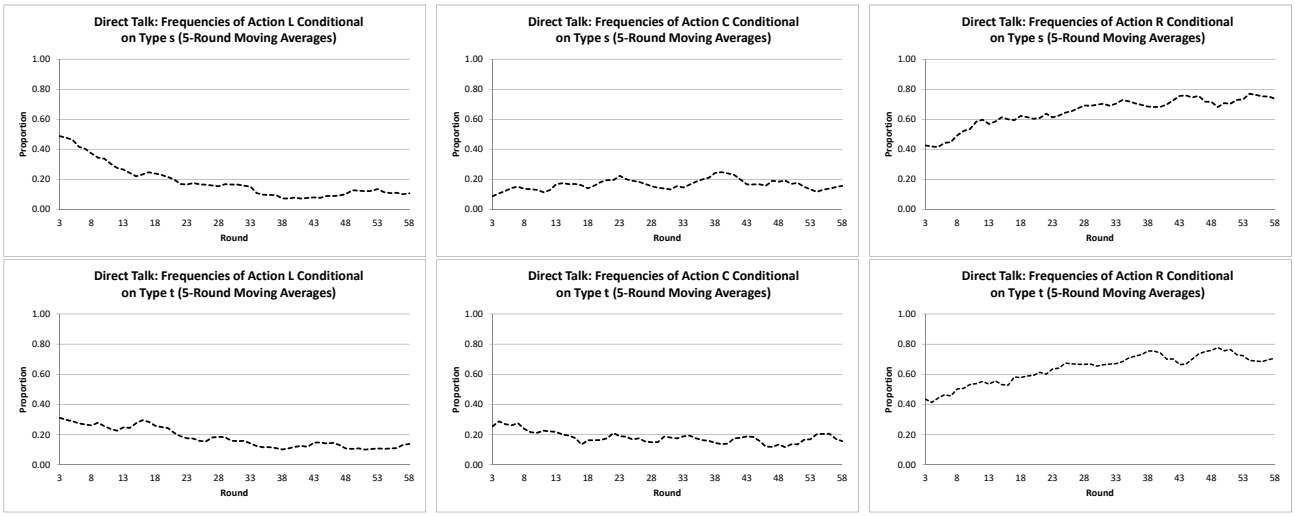
Under mediated talk, the frequency of action L conditional on type s is fairly stable at just under 40%. This is much closer to the separation prediction of 50% than pooling, which would be 0%. Departures from the separation prediction tend to be in the form of a slightly higher frequency of action R conditional on type s and a non-negligible fraction of actions C conditional on type t. As discussed in the appendix, this is consistent with a variant of our level-k analysis that allows for a portion of subjects who ignore the fact that there is mediation in the mediated-talk environment.

5.1.4 Mediated vs direct talk - payoffs

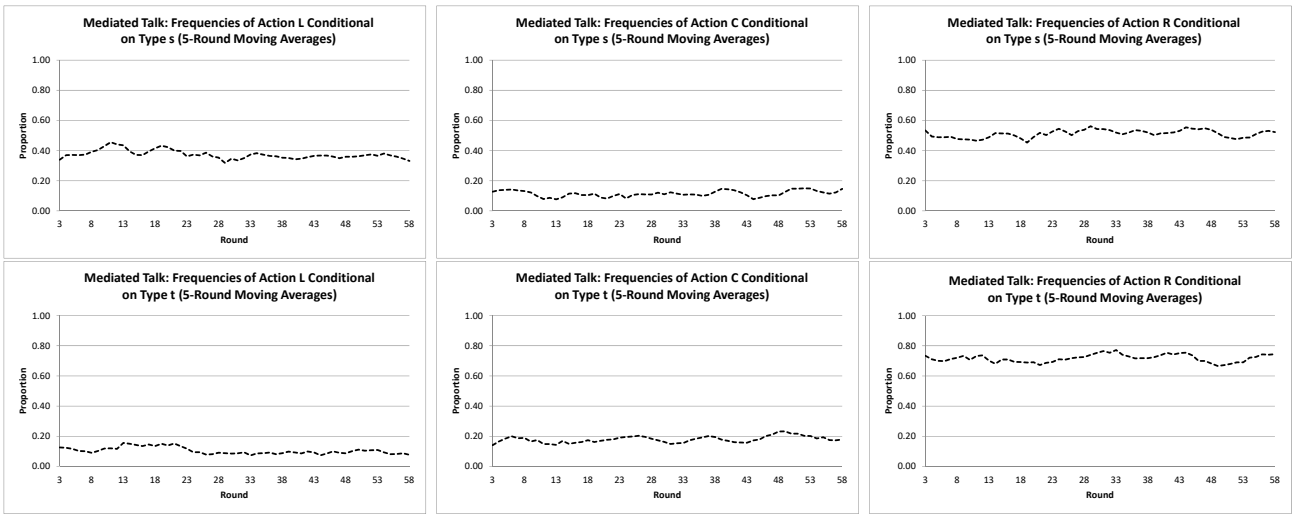
Figure 6 shows the payoffs in the mediated-talk and direct-talk treatments over the first and last ten rounds, and compares them to predicted behavior. The prediction is that (expected) payoffs for both senders and receivers are higher under mediated talk. Consistent with this prediction, we observe higher payoffs for senders as well as receivers with mediation both in the first and last 10 rounds. In the first 10 rounds under mediated talk, senders and receivers receive, respectively, average payoffs of 89.22 and 90.03, significantly higher than their respective average payoffs of 80.46 and 84.66 under direct talk ($p < 0.01$ Mann-Whitney tests); in the last 10 rounds under mediated talk, senders and receivers receive, respectively, 85.75 and 90.44, significantly higher than their respective average payoffs of 80.87 and 85.62 under direct talk ($p < 0.05$ Mann-Whitney tests). It is also worth noting that receivers do not achieve the pooling payoff with either direct or mediated talk, even though they could guarantee that payoff by simply ignoring messages. In sum, while players do not manage to approximate the full benefits from mediation, there appears to be a moderate payoff advantage from mediated talk over direct talk.

5.2 Variation in and determinants of individual behavior

In this section we explore how well individual behavior can be accounted for by our level-k model and how individual behavior is affected by information, behavioral considerations, treatment,



(a) Direct-Talk Treatments



(b) Mediated-Talk Treatments

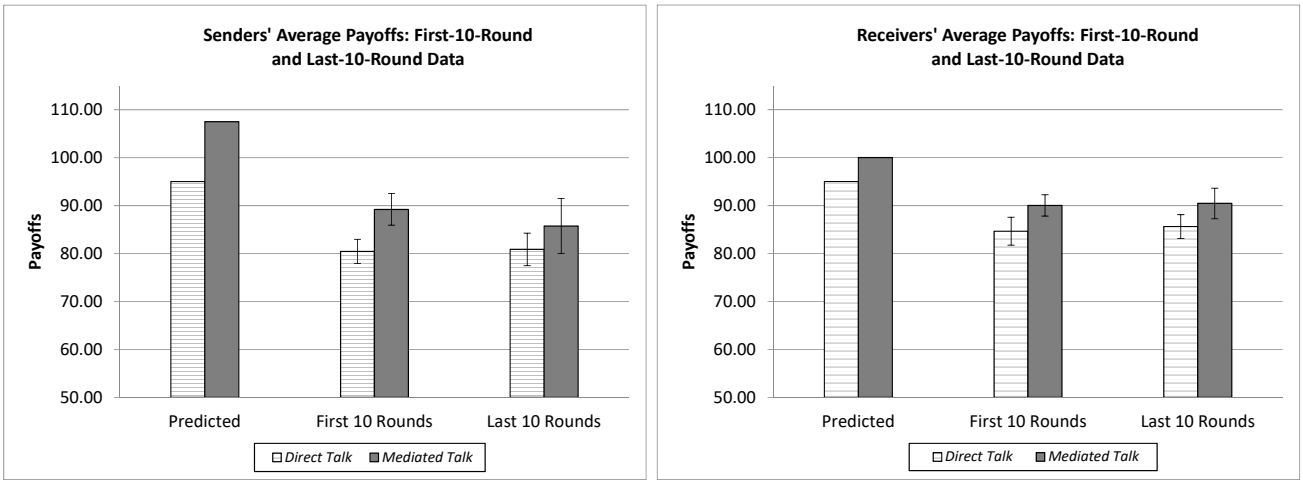
Figure 5: Trends of Frequencies of Actions Conditional on Types (5-Round Moving Averages)

prior experience, and interactions of these factors.

5.2.1 Level-k classification

For each of the treatments, our level-k model makes a single prediction for levels 1 and above. As we have seen, there are substantial departures from that prediction. To get a clearer picture of the nature of these departures, here we classify individual subjects according to whether their behavior is best described as level-0, level-k with $k \geq 1$, or resists classification.

In the case of direct talk, for both senders and receivers, strategies are different for L_0 and $L_{k\geq 1}$. For each subject, we calculate the frequencies of observed choices that are consistent with a given level for that subject's role. Note that L_0 and $L_{k\geq 1}$ strategies share some common



(a) Senders' Average Payoffs (b) Receivers' Average Payoffs

Figure 6: Average Payoffs in Direct Talk vs. Mediated Talk: First-10-Round and Last-10-Round Data

Table 13: Proportion of L_0, $L_{k\geq 1}$, and Unclassified: Sender-Subjects

	Direct Declaratives			Direct Directives		
Session	L_0	$L_{k\geq 1}$	Unclassified	L_0	$L_{k\geq 1}$	Unclassified
1	0.20	0.60	0.20	0.00	1.00	0.00
2	0.00	0.80	0.20	0.10	0.80	0.10
3	0.14	0.57	0.29	0.11	0.89	0.10
4	0.00	1.00	0.00	0.00	0.80	0.20
5	0.25	0.75	0.00	0.00	1.00	0.00
Mean	0.12	0.74	0.14	0.04	0.90	0.06

Note:

Table 14: Proportion of L_0, $L_{k\geq 1}$, and Unclassified: Receiver-Subjects

	Direct Declaratives			Direct Directives		
Session	L_0	$L_{k\geq 1}$	Unclassified	L_0	$L_{k\geq 1}$	Unclassified
1	0.00	0.50	0.50	0.00	0.50	0.50
2	0.00	0.60	0.40	0.00	0.60	0.40
3	0.00	0.71	0.29	0.00	0.33	0.67
4	0.00	0.50	0.50	0.00	0.30	0.70
5	0.00	0.38	0.62	0.00	0.50	0.50
Mean	0.00	0.54	0.46	0.00	0.45	0.55

Note:

components. This is the case when the sender's type is s or when the receiver observes message "s/L". Observed choices that are consistent with both L_0 and $L_{k\geq 1}$ classifications, are counted toward the frequencies of both L_0 and $L_{k\geq 1}$. For each subject, there will be one frequency for L_0 and another for $L_{k\geq 1}$. We single out the more frequent one. If this higher frequency is no less than 70%, the subject is classified as belonging to that level. Otherwise, the subject is considered unclassified. For each session and each role, we calculate the proportions of L_0-subjects, $L_{k\geq 1}$-

subjects, and unclassified subjects. Tables 13 and 14 present the findings for, respectively, senders and receivers.

The classification is imperfect, with the degree of conformity with the level-k prediction varying both across sessions and across treatments. One characteristic of the classification that stands out is that fewer receiver subjects than sender subjects can be classified by our rule. On average, around 50% of receiver subjects cannot be classified as using a level-k strategy, whereas more for receivers more than 80% can be classified. In both cases, when subjects can be classified, they are overwhelmingly categorized as level $L_{k\geq 1}$ rather than level L_0 players, consistent with the notion that the level zero type is only a mental construct, the model used by the lowest level "real" type.

With mediated talk, for both senders and receivers, the strategies are the same at all levels. The classification is therefore dichotomous: subjects are either classified as $L_{k\geq 0}$ players or remain unclassified.

Table 15: Proportion of $L_{k\geq 0}$ and Unclassified: Sender-Subjects

	Mediated Declaratives		Mediated Directives		Mediated Direct Mechanism	
Session	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified
1	0.90	0.10	0.89	0.11	0.90	0.10
2	0.75	0.25	0.90	0.10	1.00	0.00
3	1.00	0.00	0.80	0.20	1.00	0.00
4	0.89	0.11	0.56	0.44	0.29	0.71
5	0.70	0.30	1.00	0.00	0.78	0.22
Mean	0.85	0.15	0.83	0.17	0.79	0.21

Note:

Table 16: Proportion of $L_{k\geq 0}$ and Unclassified: Receiver-Subjects

	Mediated Declaratives		Mediated Directives		Mediated Direct Mechanism	
Session	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified
1	0.60	0.40	0.67	0.33	0.80	0.20
2	0.50	0.50	0.60	0.40	1.00	0.00
3	0.67	0.33	0.50	0.50	0.67	0.33
4	0.78	0.22	0.33	0.67	0.29	0.71
5	0.40	0.60	1.00	0.00	0.44	0.56
Mean	0.59	0.41	0.62	0.38	0.64	0.36

Note:

We use the same 70% threshold. For each subject, we calculate the frequency of observed choices that are consistent with the $L_{k\geq 0}$ strategy of his/her role. If this frequency is no less than 70%, the subject is classified as a $L_{k\geq 0}$-type. Otherwise, he/she is considered unclassified. Tables 15 and 16 present the findings for, respectively, senders and receivers. As in the case of direct talk there is considerable variation across sessions and treatments, and senders are more frequently classified as $L_{k\geq 0}$ players than are receivers.

5.2.2 Determinants of individual behavior

In order to better understand sender behavior, we regress $\mathbb{I}\{m_{i,\tau} = \text{“}s/L\text{”}\}$, an indicator variable that takes the value 1 (and zero otherwise) if sender i sends message “s/L” in period τ, on a combination of variables that are either suggested by theory, motivated by behavioral considerations, or have a plausible role in learning. In addition to a constant, our regressors are:

$\mathbb{I}\{\theta_{i,\tau} = s\}$: This indicator variable takes on the value 1 if sender i's type in period τ is s. We expect the coefficient associated with this regressor to be positive if there is some preference for truth-telling, or if some of the senders are level-0 players.

$\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{M\}$: This product of indicators takes on the value 1 (and zero otherwise) if the sender's type is s and the treatment is M (which stands for mediated talk). Theory, that is the level-k analysis for levels $k \geq 1$ and the equilibrium selected by that level-k analysis, predicts that the sender having type s makes no difference for the probability of message “s/L” being sent under direct talk and increases that probability under mediated talk. Accordingly, we expect the coefficient of the regressor $\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{M\}$ to be positive; this is the principal predicted treatment effect for senders.

$\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\}$: $\hat{m}_{i,\tau-1}$ is the message observed by the receiver that was matched with sender i in the previous period (after mediation in the mediated talk treatments) and action L is the preferred action of a sender with type s. Therefore, if the product of indicators $\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\}$ equals 1, a sender whose type is currently s is confronted with a situation where having sent message “s/L” and having that message be observed by the receiver resulted in a success in the prior period. If we believe the sender to be *positively influenced by success*, we expect the coefficient on that regressor to be positive.

$\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = C\}$: Action C is the least preferred action of both types of the sender. Therefore if in the previous period sender i effectively sent message “t/R” and the receiver's response was to take action C, she might be dissuaded from sending message “t/R” and therefore be more likely to send message “s/L. If senders are responsive to success and failure and capable for *counterfactual reasoning*, we expect the coefficient of this regressor to be positive.

$\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\} \times \mathbb{I}\{M\}$: This regressor interacts our measure of past success with message “s/L” with the treatment indicator. If there is a fraction of senders who have truth-telling preferences at the outset or, equivalently, act as level-0 players this creates a tension with the level-k prediction for levels $k \geq 1$ and our equilibrium prediction for direct talk, but not for mediated talk. If learning plays a role in resolving that

tension, then we expect learning to play a greater role under direct talk. This would suggest that the coefficient on the regressor $\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\} \times \mathbb{I}\{M\}$ is negative.

$\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = C\} \times \mathbb{I}\{M\}$: This regressor interacts our measure of past failure with sending the alternative message “t/R” with the treatment indicator. By the same reasoning as for the previous case, we expect learning to play a greater role with direct talk. Thus a negative experience with sending message “t/R” ought to have a lesser positive effect on the probability of sending message “s/L” under mediated talk. This would suggest that the coefficient on the regressor $\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = C\} \times \mathbb{I}\{M\}$ is negative.

The final two regressors examine the impact of time on how treatment affects learning. We expect the lesser role of learning under mediated talk to be more pronounced early, before behavior has settled down. Accordingly, we expect the coefficients on $\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$ and $\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = C\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$ to be negative.

Table 17 reports the estimation results for senders with standard errors clustered at the session level. We estimate both linear probability models (with fixed and random effects) and probit models (with random effects) using panel data.[9] The main takeaway from the estimations is that there is a strong treatment effect: the coefficient on $\mathbb{I}\{\theta_{i,\tau} = s\}$ has the expected sign, is large in size, and highly significant across all our specifications. In all specifications that include interaction effects with the treatment ($\mathbb{I}\{M\}$), the coefficient on $\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\}$ has the expected sign and is statistically significant, suggesting that learning from success plays a role. There is also some indication that this effect is confined to direct talk: The coefficient on $\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\} \times \mathbb{I}\{M\}$ is negative in all three specifications where it is included and highly significant in two out of three specifications. The sizes of the learning effects uniformly are an order of magnitude smaller than the treatment effect.

[9] It is likely that there are correlations between unobserved heterogeneity and some of the regressors, which makes fixed effects more preferred. For example, sender-subjects may differ in their expectations of how receiver-subjects would respond as a result of the messages they sent. These expectations are likely to correlate with the communication environment (direct or mediated talk), the message delivered in the prior round, and the action taken in the prior round. A Hausman test further confirms that random effects are not preferred ($p < 0.001$). For the probit model we only consider random effects because pairing probit with fixed effects is problematic.

Table 17: Linear Probability and Probit Models: Senders (Standard Errors Clustered at Session Level)

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.4636***	0.4648***	–	0.4637***	0.4653***	–
	(0.0127)	(0.0656)	–	(0.0128)	(0.0657)	–
$\mathbb{I}\{\theta_{i,\tau} = s\}$	0.0621*	0.1128***	0.0579**	0.0540*	0.1165***	0.0547*
	(0.0249)	(0.0256)	(0.0217)	(0.0256)	(0.0265)	(0.0216)
$\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{M\}$	0.6877***	0.6017***	0.5095***	0.7011***	0.5955***	0.5141***
	(0.0451)	(0.0392)	(0.0301)	(0.0456)	(0.0385)	(0.0301)
$\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\}$	0.0072	0.0085	0.0653***	0.0508**	0.0567***	0.0921***
	(0.0130)	(0.0124)	(0.0179)	(0.0183)	(0.0177)	(0.0272)
$\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = C\}$	0.0050	−0.0044	0.0022	−0.0549	−0.0308	−0.0272
	(0.0138)	(0.0149)	(0.0138)	(0.0297)	(0.0294)	(0.0172)
$\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\} \times \mathbb{I}\{M\}$	–	–	–	−0.0575*	−0.0655**	−0.0244
	–	–	–	(0.0249)	(0.0238)	(0.0461)
$\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = C\} \times \mathbb{I}\{M\}$	–	–	–	0.0777*	0.0290	0.0559*
	–	–	–	(0.0316)	(0.0300)	(0.0231)
$\mathbb{I}\{\theta_{i,\tau} = s\} \times \mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = L\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$	–	–	–	−0.0324	−0.0306*	−0.0832
	–	–	–	(0.0167)	(0.0155)	(0.0480)
$\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = C\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$	–	–	–	0.0031	0.0036	−0.0186
	–	–	–	(0.0179)	(0.0176)	(0.0236)
No. of Observations	13,452	13,452	13,452	13,452	13,452	13,452

Note: Column (1) reports the coefficients from estimating the fixed-effects linear probability model with four explanatory variables. Column (2) reports the coefficients from estimating the random-effects linear probability model with four explanatory variables. Column (3) reports the average marginal effects from estimating the random-effects probit model with four explanatory variables. Column (4) reports the coefficients from estimating the fixed-effects linear probability model with eight explanatory variables. Column (5) reports the coefficients from estimating the random-effects linear probability model with eight explanatory variables. Column (6) reports the average marginal effects from estimating the random-effects probit model with eight explanatory variables. Standard errors clustered at the subject level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

Our examination of receiver behavior parallels that for senders. We regress $\mathbb{I}\{a_{j,\tau} = L\}$, an indicator variable that takes the value 1 (and zero otherwise) if receiver j takes action L in period τ, on a combination of variables that are either suggested by theory, motivated by behavioral considerations, or have a plausible role in learning. In addition to a constant, our regressors are:

$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\}$: This indicator variable takes on the value 1 if receiver j observes message “s/L” in period τ. We expect the coefficient associated with this regressor to be positive if there is some fraction of credulous receivers.

$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\}\times\mathbb{I}\{M\}$: This product of indicators takes on the value 1 (and zero otherwise) if receiver j observes message “s/L” in period τ and the treatment is M (mediated talk). Theory, that is the level-k analysis for levels $k \geq 1$ and the equilibrium selected by that level-k analysis, predicts that observing message “s/L” induces action L under mediated but not under direct talk. Therefore, we expect the coefficient of the regressor $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{M\}$ to be positive; this is the principal predicted treatment effect for receivers.

$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = s\}$: The message $\tilde{m}_{j,\tau-1}$ is the message that was sent by the sender who was matched with receiver j in period $\tau - 1$ (prior to possible garbling by mediation). If the current message is “s/L” and in the prior period that message was sent by type s, we might expect a receiver who learns to be more inclined to respond with L, the action that is a best response to beliefs concentrated on type s. Therefore the coefficient on the regressor $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = s\}$ ought to be positive.

$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\}\times\mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\}\times\mathbb{I}\{\theta_{j,\tau-1} = t\}$: If the current message is “s/L” and in the prior period the alternative message “t/R” was sent by type t, we might expect a receiver who learns and engages in counterfactual reasoning to be more inclined to respond with action L, the action that is a best response to beliefs concentrated on type s. Therefore the coefficient on the regressor $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = t\}$ also ought to be positive.

With mediated talk, for the reasons given above, learning might be playing less of a role. If so, this would diminish the direct positive impact the probability of taking action L from seeing message “s/L” associated with type s as well as the indirect positive impact on that probability from seeing message “t/R” associated with type t. Then we would expect the coefficients on $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = s\} \times \mathbb{I}\{M\}$ and on $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = t\} \times \mathbb{I}\{M\}$ to be negative. In the case of receivers there is, however, a potential countervailing effect: they might find it more difficult to associate messages with types under mediation, which would increase the role of learning, and render the coefficients positive.

The final two regressors examine the impact of time on how treatment affects learning. We expect the role of learning under mediated talk to be more pronounced early, before behavior has settled down. Accordingly, if learning is less important with mediation, we expect the coefficients on $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = s\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$ and $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = t\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$ to be negative. If, however, learning is more important for receivers under mediation, we expect those coefficients to be positive.

Table 18 reports the estimation results for receivers with standard errors clustered at the session level. Similar to the estimations for senders, we estimate a linear probability model with both random and fixed effects and a probit model with random effects. Again, the treatment effect, as measured by the coefficient on $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{M\}$, is highly significant in all of our specifications and dominates all other effects in size. In addition, we see that credulousness of the receiver, as measured by the coefficient on $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\}$, plays an important role; the coefficient estimate is smaller than for the treatment effect, but not by orders of magnitude. Learning appears to play a greater role for receivers than for senders. There is a highly significant effect of learning from counterfactuals, as measured by the coefficient on $\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = t\}$ in all of our specification. It is of a similar size as the credulousness effect. In the first 20 rounds, where we expect most learning to take place, receiver behavior appears to be more impacted by learning under mediated talk than under direct talk.

Table 18: Linear Probability and Probit Models: Receivers (Standard Errors Clustered at Session Level)

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.0424**	0.0398***	–	0.0427**	0.0400***	–
	(0.0148)	(0.0080)	–	(0.0147)	(0.0080)	–
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\}$	0.1112***	0.1192***	0.1145***	0.1162***	0.1246***	0.1145***
	(0.0284)	(0.0247)	(0.0278)	(0.0271)	(0.0238)	(0.0272)
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{M\}$	0.5237***	0.5123***	0.3043***	0.4809***	0.4683***	0.2992***
	(0.0674)	(0.0647)	(0.0243)	(0.0674)	(0.0665)	(0.0220)
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = s\}$	0.0621**	0.0634***	0.0326**	0.0412	0.0422*	0.0264*
	(0.0185)	(0.0186)	(0.0105)	(0.0211)	(0.0212)	(0.0125)
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = t\}$	0.1326***	0.1340***	0.0740***	0.1683***	0.1697***	0.0917***
	(0.0213)	(0.0217)	(0.0139)	(0.0352)	(0.0356)	(0.0171)
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = s\} \times \mathbb{I}\{M\}$	–	–	–	0.0520	0.0531	0.0035
	–	–	–	(0.0356)	(0.0362)	(0.0219)
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = t\} \times \mathbb{I}\{M\}$	–	–	–	−0.0533	−0.0526	−0.0508*
	–	–	–	(0.0456)	(0.0469)	(0.0236)
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = s\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$	–	–	–	0.1083*	0.1090**	0.0609*
	–	–	–	(0.0393)	(0.0392)	(0.0243)
$\mathbb{I}\{m_{j,\tau} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\tilde{m}_{j,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{\theta_{j,\tau-1} = t\} \times \mathbb{I}\{M\} \times \mathbb{I}\{\tau \leq 20\}$	–	–	–	0.1102*	0.1099***	0.0773*
	–	–	–	(0.0433)	(0.0433)	(0.0322)
No. of Observations	13,452	13,452	13,452	13,452	13,452	13,452

Note: Column (1) reports the coefficients from estimating the fixed-effects linear probability model with four explanatory variables. Column (2) reports the coefficients from estimating the random-effects linear probability model with four explanatory variables. Column (3) reports the average marginal effects from estimating the random-effects probit model with four explanatory variables. Column (4) reports the coefficients from estimating the fixed-effects linear probability model with eight explanatory variables. Column (5) reports the coefficients from estimating the random-effects linear probability model with eight explanatory variables. Column (6) reports the average marginal effects from estimating the random-effects probit model with eight explanatory variables. Standard errors clustered at the subject level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

6 Discussion

In this paper we conduct a communication design exercise, trying to see whether mediation can improve information transmission. We find a definite treatment effect. There are pronounced differences in behavior and outcomes between direct and mediated communication. Direct communication tends toward pooling, whereas mediated communication tends toward separation. This translates into moderate payoff improvements for both senders and receivers from replacing direct with mediated communication.

Theory, which predicts pooling under direct talk, separation under mediated talk, and the specific way in which messages are used, captures modal behavior well. There are, however, substantial and sometimes stable departures from the theoretical prediction. Confirming results from prior work on communication games, we find over-communication by senders in the initial rounds of direct talk; this over-communication vanishes with experience. In contrast, there is stable under-communication by senders under mediated talk. Receivers over-interpret messages under both direct and mediated talk, i.e., with non-negligible frequency they take actions that would only be optimal if they had more precise information. This is particularly striking under mediated communication, where a sizable fraction of receiver subjects use strictly dominated strategies.

The challenge for future work will be to see whether the lessons learned from our deliberately chosen simple setting extend to other communication environments. Specifically, it would be nice to know whether there is a robust mediation scheme that facilitates communication for a broad range of incentives and with relaxed knowledge requirements about the nature of private information.

References

[1] Basu, Kaushik, and Jörgen W. Weibull [1991], "Strategy Subsets Closed under Rational Behavior," *Economics Letters* **36**, 141-146.

[2] Blume, Andreas, Oliver J. Board, and Kohei Kawamura [2007], "Noisy Talk," *Theoretical Economics* **2**, 395-440.

[3] Blume, Andreas, Ernest K. Lai, and Wooyoung Lim [2017], "Strategic Information Transmission: A Survey of Experiments and Theoretical Foundations," Working Paper.

[4] Blume, Andreas, Ernest K. Lai, and Wooyoung Lim [2018], "Eliciting Private Information with Noise: The Case of Randomized Response," *Games and Economic Behavior*, forthcoming.

[5] Cai, Hongbin, and Joseph Tao-Yi Wang [2006] "Overcommunication in Strategic Information Transmission Games," *Games and Economic Behavior* **56**, 7-36.

[6] Crawford, Vincent P. and Joel Sobel [1982], "Strategic Information Transmission," *Econometrica* **50**, 1431–1451.

[7] Crawford, Vincent P. [2003], "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions," *American Economic Review* **93**, 133-149.

[8] Fischbacher, Urs [2007], "z-tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics* **10**, 171–178.

[9] Forges, Françoise [1985], "Correlated Equilibria in a Class of Repeated Games with Incomplete Information," *International Journal of Game Theory* **14**, 129–149.

[10] Fréchette, Guillaume, Alessandro Lizzeri, Jacopo Perego [2018], "Rules and Commitment in Communication," Working Paper.

[11] Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani [2009], "Mediation, Arbitration and Negotiation," *Journal of Economic Theory* **144**, 1397-1420.

[12] Kalai, Ehud, and Dov Samet [1984], "Persistent Equilibria in Strategic Games," *International Journal of Game Theory* **13**, 129-144.

[13] Kamenica, Emir, and Matthew Gentzkow [2011], "Bayesian Persuasion," *American Economic Review* **101**, 2590-2615.

[14] Krishna, Vijay, and John Morgan [2004], "The Art of Conversation: Eliciting Information from Experts through Multi-Stage Communication," *Journal of Economic Theory* **117**, 147-179.

[15] Myerson, Roger B [1982], "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," *Journal of Mathematical Economics* **10**, 67-81.

[16] Myerson, Roger B. [1991], *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, MA.

[17] Nguyen, Quyen [2016], "Bayesian Persuasion: Evidence from the Laboratory," Utah State University, Working Paper.

[18] Wang, Joseph Tao-yi, Michael Spezio, and Colin F. Camerer [2010], "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review* **100**, 984-1007.

[19] Warner, Stanley L. [1965], "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association* **60**, 63-69.

A Appendix: equilibrium analysis

For the equilibrium analysis the framing of messages as directives or declaratives does not matter: given any equilibrium under one frame, there exists an outcome-equivalent equilibrium under any alternative frame, in which the message are simply renamed.

A.1 The set of equilibrium outcomes with direct talk

The receiver has seven possible types of responses to any message received: he can randomize over all three actions, L, C, and R , randomize over two, e.g. C and R, or take a single action with probability one. Any randomization with L and C in the support is ruled out as a best response, because whenever the receiver is indifferent between L and C he strictly prefers R. In any pairwise comparison of the remaining five response types, the sender in state s strictly prefers one of them. It follows that if two messages are sent with positive probability in equilibrium and the receiver responds differently after those message, the sender in state s will have a strict preference for one of the messages, leaving the other message sent exclusively in state t. This, however, cannot constitute an equilibrium, because the message that is exclusively sent in state t identifies the state and leads to the least preferred action for the sender. Therefore, in any equilibrium either only one message is sent or the receiver's responses do not vary with the messages. In both cases, the receiver takes action R in response to any message that is sent in equilibrium. It follows that ***pooling is the only equilibrium outcome with direct communication***. There are many equilibria that support the pooling outcome. The equilibrium analysis does not discriminate among these equilibria, although it seems reasonable to expect that the framing of messages impacts behavior.

A.2 The set of equilibrium outcomes with mediated talk

Since the framing of messages is irrelevant for the equilibrium analysis, we will take advantage of the notational convenience of conducting this analysis in the specific framework of mediated declaratives. Recall that with mediated declaratives the set of sent messages $\{s, t\}$ coincides with the set of received messages. Pooling, where the receiver responds to all messages received in equilibrium with action R, is supported by many equilibria. Suppose instead that the receiver observes both messages with positive probability in equilibrium, and responds differently to different messages. Call such an equilibrium *influential*. Since the receiver's expected posterior equals the prior, he will mix (possibly degenerately) over L and R after one message and over C and R after the other. We will refer to the former type of lottery by LR and the latter by CR.

Regarding influential equilibria, there are two cases to consider:

Case 1: The receiver uses LR after message s.

By sending message s the sender induces LR with probability one half and CR otherwise. By sending message t she induces CR with probability one. Since at least one of the two lotteries assigns positive probability to an action other than R, type s strictly prefers sending message s. For the receiver to treat messages s and t differently type t must send message t with probability greater than zero. Suppose type t sends message t with probability one. Then the receiver assigns posterior probability $\frac{2}{3}$ to type t after receiving message t. Thus action R is the unique best reply after receiving message t. Therefore we have an equilibrium in pure strategies in which sender type s sends message s, sender type t sends message t, the receiver responds to message s with action L and to message t with action R.

Suppose type t sends message t with probability $x \in (0,1)$. Then the posterior probability of type t given message t is

$$\gamma := \frac{x\frac{1}{2}}{x\frac{1}{2} + \sigma(s|s)\frac{1}{2}\frac{1}{2}} = \frac{2x}{2x+1}.$$

The posterior probability of type t given message s is

$$\delta := \frac{(1-x)\frac{1}{2}\frac{1}{2}}{(1-x)\frac{1}{2}\frac{1}{2} + \sigma(s|s)\frac{1}{2}\frac{1}{2}} = \frac{1-x}{2-x}.$$

For the receiver to be indifferent between C and R following message t would require that $120\gamma = 90\gamma + 100(1-\gamma)$, or $\gamma = \frac{10}{13}$. For this we would need $x = \frac{5}{3}$, which is impossible. For the receiver to be indifferent between L and R following message s would require that $120(1-\delta) = 90\delta + 100(1-\delta)$, or $\gamma = \frac{2}{11}$. For this we need $x = \frac{7}{9}$. With $x = \frac{7}{9}$ the receiver takes action R with probability one after receiving message t. This gives the sender a choice between inducing an LR lottery that assigns probability less than one to R and inducing R with probability one. This implies that sender type t strictly prefers sending message t, which contradicts $x \in (0,1)$.

Thus for the case under consideration, the only influential equilibrium is the one in in which type s sends message s and type t sends message t, which supports the outcome we refer to as *separation*.

Case 2: The receiver uses LR after message t.

By sending message t, the sender induces LR with probability one. By sending message s the sender induces LR with probability one half and CR otherwise. Since at least one of the two lotteries assigns probability less than one to action R, type s of the sender strictly prefers sending message t. For the receiver to respond differently after the two messages, type t must send message s with probability greater than zero. Then, since only type t sends message s with

positive probability the receiver's unique best reply to message s is action C, which results in sender type t receiving her lowest possible payoff. Since the case assumes that the receiver mixes between L and R following message t, type t has a strict preference for message t over message s. This implies that the two types pool on message t contradicting our assumption that we have an influential equilibrium. Thus there is no influential equilibrium in this case.

In summary, since an equilibrium that is not influential is a pooling equilibrium, we have shown that ***with mediation separation and pooling are the only two equilibrium outcomes.*** Furthermore, we have found that ***there is a unique equilibrium supporting separation***, which implies that ***for the case of separation, the equilibrium analysis pins down message use.*** For the case of pooling, the equilibrium analysis does not pin down message use.

B Appendix: level-k analysis of mediation with a fraction λ of players who ignore mediation

Table 19: Level-k Prediction for the Direct-Declaratives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	C
$L_{k\geq 1}$	"s"	"s"	R	C

Table 20: Level-k Prediction for the Mediated-Declaratives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	R
$L_{k\geq 1}$	"s"	"t"	L	R

Tables 19 and 20 reproduce the level-k analysis of the two declaratives games for easy reference. The analysis below is conducted in terms of declaratives, but is the same for directives and the mediated direct mechanism, modulo adjusting the notation.

Suppose there is a fraction $\lambda \in (0,1)$ of players who ignore the fact that there is mediation in the mediated-talk treatments. The behavior of these players is pinned down by the level-k analysis of the direct declaratives game, reproduced in Table 19. For the remaining fraction $1-\lambda$ of players who understand that there is mediation, we have two options: they either recognize that there is a fraction λ of players who ignore mediation, or they do not recognize this fact. In the latter case, where they do not recognize, they analyze the game as indicated in Table 20. If instead they do recognize, we can ask how small λ has to be for the analysis to remain the same as in Table 20.

Consider receivers first. For L_0 receivers it does not matter whether L_0 senders do or do not ignore mediation. $L_{k\geq 1}$ receivers who recognize that a fraction λ of senders ignore mediation assign posterior probability $\frac{1}{1+\lambda}$ to type s after observing message "s." For action L to remain optimal after message "s" requires $\frac{1}{1+\lambda}120 \geq \frac{1}{1+\lambda}100 + \frac{\lambda}{1+\lambda}90$, which is satisfied if $\lambda \leq \frac{2}{9}$. L_1 receivers who recognize that a fraction λ of senders ignore mediation assign posterior probability $\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}\lambda+(1-\lambda)} = \frac{1}{3-\lambda}$ to type s after observing message "t". For action R to remain optimal after message "t" imposes no additional constraint on λ.

Consider senders next. L_0 sender behavior is pinned down by assumption. Thus, it suffices to check incentives for $L_{k\geq 1}$ senders. If optimality is maintained for $L_{k\geq 2}$ senders, then it is also maintained for L_1 senders. An $L_{k\geq 2}$ sender's payoff if her type is s and she sends message "s" equals $(1-\lambda)\left(\frac{1}{2}110 + \frac{1}{2}60\right) + \lambda\left(\frac{1}{2}60 + \frac{1}{2}10\right)$. If instead she sends message "t" her payoff is

$(1-\lambda)60+\lambda 10$. Thus type s prefers sending message "s" regardless of the value of λ. An $L_{k\geq 2}$ sender's payoff if her type is t and she sends message "t" equals $(1-\lambda)130+\lambda 10$. If instead she sends message "s" her payoff is $(1-\lambda)\left(\frac{1}{2}80+\frac{1}{2}130\right)+\lambda\left(\frac{1}{2}10+\frac{1}{2}130\right)$. Thus for it to be optimal that type t sends message "t", it is necessary and sufficient that $\lambda \leq \frac{5}{17}$.

Combining the constraints for senders and receivers, it follows that the level-k analysis of mediated talk survives as long as $\lambda \leq \frac{2}{9}$. Similarly, in a game where a fraction $\lambda \leq \frac{2}{9}$ of players play consistent with an equilibrium of the direct-talk game and the remaining players both recognize that there is mediation and the fact that a fraction λ of players does not understand this, it is an equilibrium for the latter kind of players to use separating strategies.

C Appendix: receiver-anchored level-k analysis

Table 21: Level-k Prediction for the Direct-Declaratives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"s"	L	C
$L_{k\geq 1}$	"s"	"s"	R	C

Table 22: Level-k Prediction for the Direct-Directives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"L"	"R"	L	R
L_1	"L"	"L"	L	C
$L_{k\geq 1}$	"L"	"L"	R	C

Table 23: Level-k Prediction for the Mediated-Declaratives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	R
$L_{k\geq 1}$	"s"	"t"	L	R

Table 24: Level-k Prediction for the Mediated-Directives Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"L"	"R"	L	R
$L_{k\geq 1}$	"L"	"R"	L	R

Table 25: Level-k Prediction for the Mediated-Direct-Mechanism Game

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	R
$L_{k\geq 1}$	"s"	"t"	L	R

In the receiver-anchored level-k analysis level-0 (L_0) receivers are assumed to be credulous in response to declaratives and obedient in response to directives. L_k senders best respond to L_k receivers for all $k \geq 0$ and L_k receivers best respond to L_{k-1} senders for all $k \geq 1$. Predictions from the receiver-anchored level-k analysis are identical to those from the sender-anchored level-k analysis for all the mediated-talk games. With direct talk there are small differences: The receiver-anchored analysis predicts that in the direct-declaratives game, L_0 senders pool on "s" and in the direct-directives game L_1 receivers respond with L to "L" and with C to "R". The Sender- and receiver-anchored analyses make the same predictions for all levels $k > 1$.

D Appendix: Individual treatments

In this section we disaggregate and report data separately for each individual treatment.

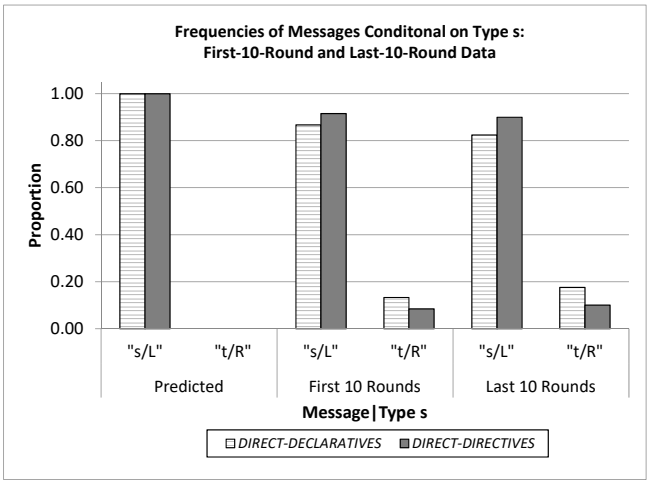
D.1 Individual treatments: senders

Figure 7 shows sender behavior in each of the five treatments, in both the first 10 rounds and the last 10 rounds, and relates observations to predictions.

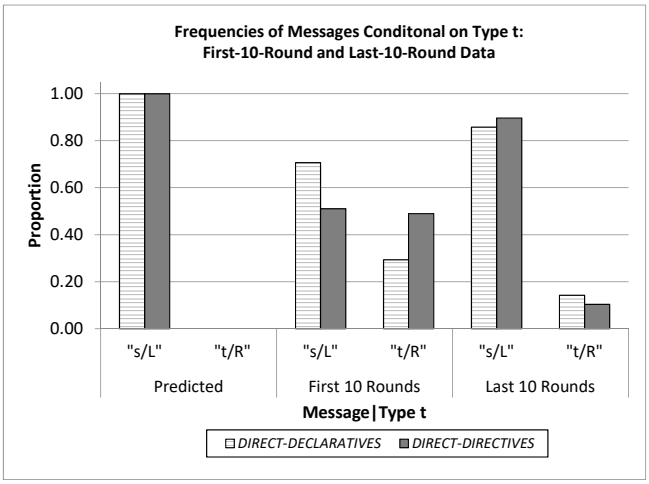
The top panels report sender behavior in each of the two direct-talk treatments. There is a noticeable difference between the two treatments in the first 10 rounds, which tends to vanish in the last 10 rounds, especially for type-t behavior. Type-s senders in the first 10 rounds send message "L" 92% of the time in the *Direct-Directives* treatment, more frequently than the 87% of message "s" in the *Direct-Declaratives* treatment (p = 0.0476, Mann-Whitney test); in the last 10 rounds they send message "L" 90% of the time in the *Direct-Directives* treatment, also more frequently than the 82% of message "s" in the *Direct-Declaratives* treatment but with a slightly lower statistical significance (p = 0.0575, Mann-Whitney test). For type-t senders, in the first 10 rounds they send message "R" 49% of the time in the *Direct-Directives* treatment, more frequently than the 29% of the message "t" in the *Direct-Declaratives* treatment (p = 0.0159, Mann-Whitney test); in the last 10 rounds they send message "R" 10% of the time in the *Direct-Directives* treatment, less frequently than the 14% of message "t" in the *Direct-Declaratives* treatment but with no statistical significance (p = 0.2317, Mann-Whitney test).

The behavior of type-t senders in the initial rounds is consistent with them trying to *direct* receivers to take their favorite action R: while our sender-anchored level-k analysis, where the anchor is forthright behavior of senders at level 0, does not distinguish between the direct-directives and direct-declaratives treatments, in a receiver-anchored level-k analysis, where the anchor is credulity of receivers at level 0 (see the appendix) of the direct-directives game, at level 0 type ts separate, sending message R, whereas in the analysis of the direct-declaratives game they pool, sending message s. Overall, given the slight difference in initial behavior and no discernible difference in terminal behavior, there appears to be no loss in pooling the data from both treatments.

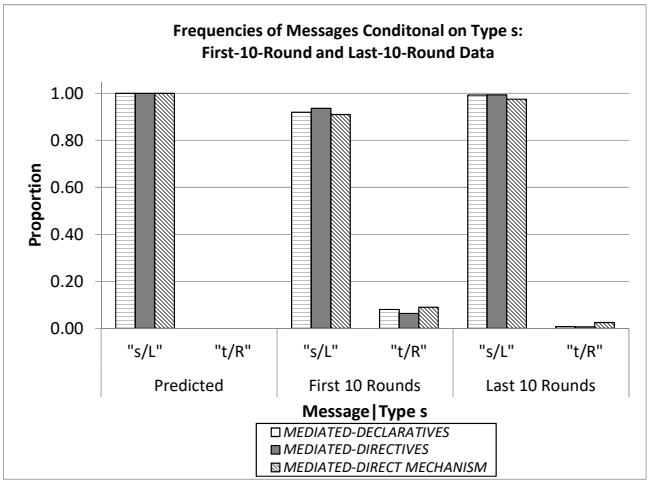
The lower panels of Figure 7 report sender behavior in each of the three mediated talk treatments. In all three treatments the modal behavior is separation from the outset, with no substantial difference across treatments. In the first 10 rounds, type-s senders send message "s" 92% of the time in the *Mediated-Declaratives* treatment, send message "L" 94% of the time in the *Mediated-Directives* treatment, and send message "s" 91% of the time in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.4206$ in any pairwise comparison, Mann-Whitney tests);



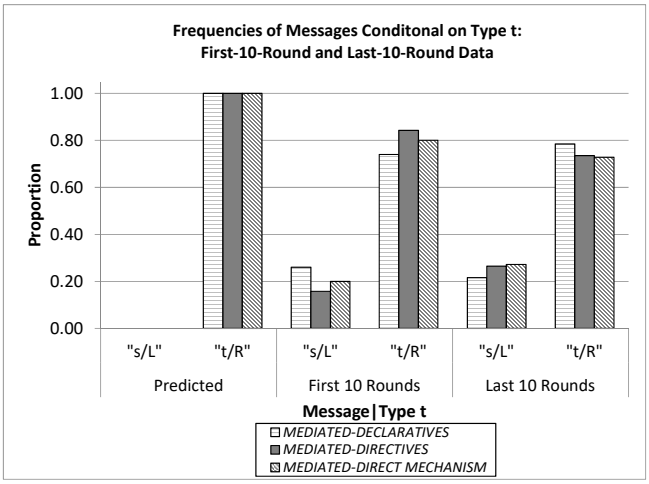
(a) Direct-Talk Treatments: Type s



(b) Direct-Talk Treatments: Type t



(c) Mediated-Talk Treatments: Type s



(d) Mediated-Talk Treatments: Type t

Figure 7: Senders' Behavior in Individual Treatments:
First-10-Round and Last-10-Round Data

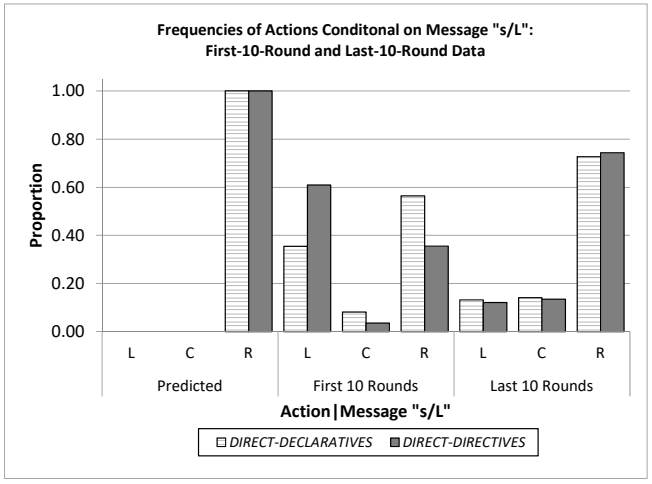
type-t senders send message "t" 74% of the time in the *Mediated-Declaratives* treatment, send message "R" 84% of the time in the *Mediated-Directives* treatment, and send message "t" 80% of the time in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.2222$ in any pairwise comparison, Mann-Whitney tests). In the last 10 rounds, type-s senders send message "s" 99% of the time in the *Mediated-Declaratives* treatment, send message "L" 99% of the time in the *Mediated-Directives* treatment, and send message "s" 97% of the time in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.2652$ in any pairwise comparison, Mann-Whitney tests); type-t senders send message "t" 78% of the time in the *Mediated-Declaratives* treatment, send message "R" 74% of the time in the *Mediated-Directives* treatment, and send message "t" 73% of the time in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.6905$ in any pairwise comparison, Mann-Whitney tests).[10] In all three treatments separation is more pronounced for

[10]For each of the four cases, a Kruskal-Wallis test further confirms that the three frequencies have no statistical

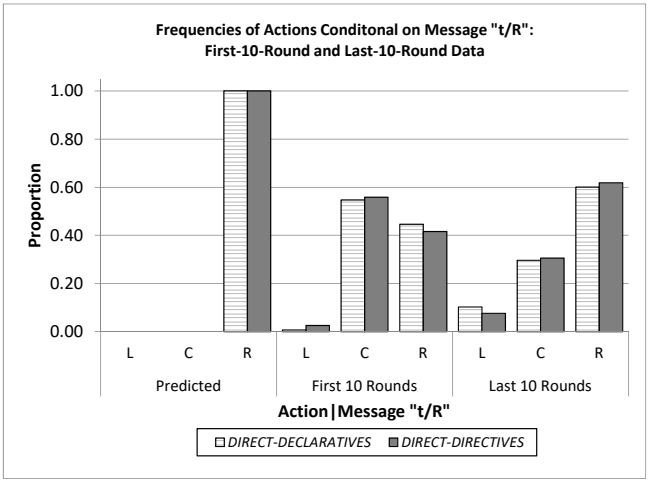
s types than for t types. Absent significant differences across treatments, it makes sense to pool the data from the three mediated-talk treatments.

D.2 Individual treatments: receivers

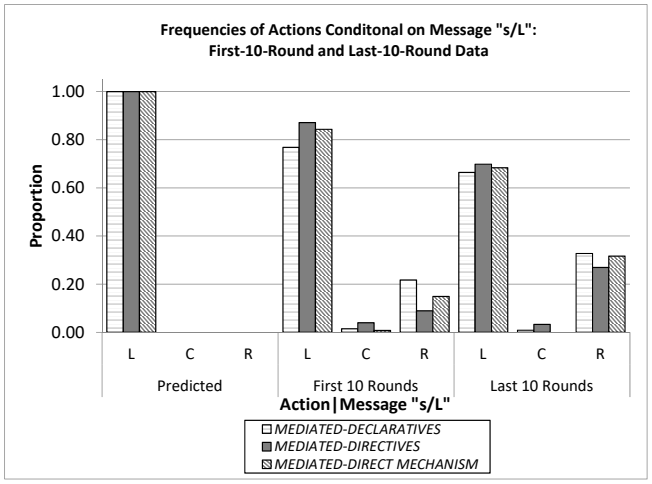
Figure 8 reports receiver behavior in each of the five treatments, in the first ten rounds and the last ten rounds.



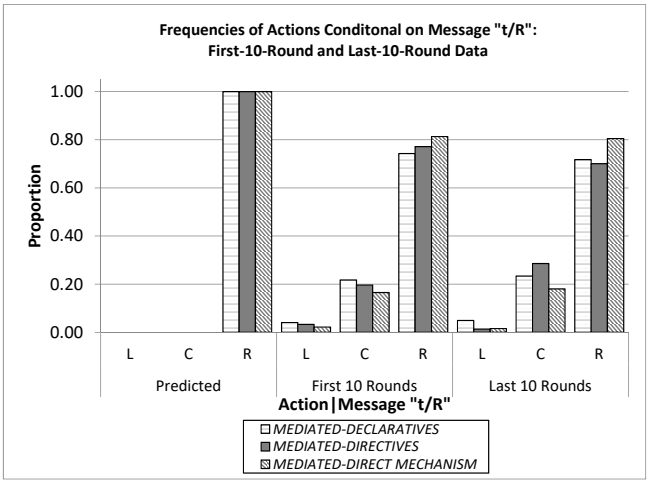
(a) Direct-Talk Treatments: Message "s/L"



(b) Direct-Talk Treatments: Message "t/R"



(c) Mediated-Talk Treatments: Message "s/L"



(d) Mediated-Talk Treatments: Message "t/R"

Figure 8: Receivers' Behavior in Individual Treatments:
First-10-Round and Last-10-Round Data

The top panels show receiver behavior in each of the two direct-talk treatments. In the last 10 rounds there is almost no difference in receiver behavior between the two direct-talk treatments. Conditional on message "s/L," the frequency of action R is 73% in the *Direct-Declaratives* treat-

differences from one another ($p \geq 0.2894$).

ment and 74% in the *Direct-Directives* treatment; conditional on message "t/R," the frequency of action R is 60% in the *Direct-Declaratives* treatment and 62% in the *Direct-Directives* treatment ($p \geq 0.8413$, Mann-Whitney tests). In the first 10 rounds, there is also no noticeable difference in receiver behavior following message "t/R" between the two treatments, but following message "s/L" receivers respond more frequently with action L in the *Direct-Directives* treatment than in the *Direct-Declaratives* treatment. Conditional on message "t/R," the frequency of action C is 55% in the *Direct-Declaratives* treatment and 56% in the *Direct-Directives* treatment (two-sided $p = 0.6905$, Mann-Whitney test); conditional on message "s/L," the frequency of action L is 61% in the *Direct-Directives* treatment, significantly higher than the 35% in the *Direct-Declaratives* treatment ($p < 0.01$, Mann-Whitney test). There is some support for this kind of initial behavior in the receiver-anchored level-k analysis: level-1 receivers respond with action L to message "L" under direct directives, whereas they respond with action R to message "s" under direct declaratives. This difference between treatments is slight and vanishes over time, suggesting that there is no loss in pooling the data from the two direct-talk treatments.

The lower panels of Figure 8 show receiver behavior in each of the three mediated-talk treatments. In all three treatments the modal behavior is separation from the outset, with no significant variations across the treatments. In the first 10 rounds, conditional on message "s/L," the frequency of action L is 77% in the *Mediated-Declaratives* treatment, 87% in the *Mediated-Directives* treatment, and 84% in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.402$ in any pairwise comparison, Mann-Whitney tests); conditional on message "t/R," the frequency of action R is 74% in the *Mediated-Declaratives* treatment, 77% in the *Mediated-Directives* treatment, and 81% in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.222$ in any pairwise comparison, Mann-Whitney tests).[11]

Departures towards action R following message "s/L" and toward action C following message "t/R" are observed in terminal behavior, which are common to all three treatments. In the last 10 rounds, conditional on message "s/L," the frequency of action R is 33% in the *Mediated-Declaratives* treatment, 27% in the *Mediated-Directives* treatment, and 32% in the *Mediated-Direct Mechanism* treatment (two-sided $p = 0.9166$ in any pairwise comparison, Mann-Whitney tests); conditional on message "t/R," the frequency of action C is 23% in the *Mediated-Declaratives* treatment, 29% in the *Mediated-Directives* treatment, and 18% in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.4206$ in any pairwise comparison, Mann-Whitney tests).[12] The homogeneity of the three mediated talk treatments suggests that pooling the data from the three treatments is without loss.

[11] For each of the two cases, a Kruskal-Wallis test further confirms that the three frequencies have no statistical differences from one another ($p \geq 0.3791$).

[12] For each of the two cases, a Kruskal-Wallis test further confirms that the three frequencies have no statistical differences from one another ($p \geq 0.6126$).

D.3 Individual treatments: outcomes

Table 26 reports the outcomes for the two direct-talk treatments over the first and last 10 rounds.

While it appears that initially type s is somewhat more readily identified in the *Direct-Directives* treatment, that difference disappears over time. In the last 10 rounds the outcomes from the two treatments are very similar, where the frequencies of the pooling action R are 70% in the *Direct-Declaratives* treatment and 74% in the *Direct-Directives* treatment (two-sided $p = 0.6905$, Mann-Whitney test).

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	15%	7%	**30%**
t	13%	10%	**25%**

First 10 Rounds

	L	C	R
s	7%	9%	**39%**
t	6%	8%	**31%**

Last 10 Rounds

(a) Direct-Declaratives

	L	C	R
s	30%	4%	**19%**
t	15%	13%	**19%**

First 10 Rounds

	L	C	R
s	6%	7%	**36%**
t	6%	8%	**37%**

Last 10 Rounds

(b) Direct-Directives

Table 26: Direct Communication Outcomes (Treatment Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

Table 27 reports the outcomes for the three mediated-talk treatments over the first and last 10 rounds. The frequencies of outcome (s, L) in the first 10 and last 10 rounds are 17% and 18% in the *Mediated-Declaratives* treatment, commonly 20% in the *Mediated-Directives* treatment, and 17% and 16% in the *Mediated-Direct Mechanism* treatment. The frequencies of outcome (t, R) in the first 10 and last 10 rounds are 36% and 32% in the *Mediated-Declaratives* treatment, 35% and 32% in the *Mediated-Directives* treatment, and 42% and 39% in the *Mediated-Direct Mechanism* treatment. The outcomes are fairly homogeneous, both across time ($p \geq 0.4227$, Wilcoxon signed-rank tests) and across treatments (two-sided $p \geq 0.1732$ in any relevant pairwise

comparison, Mann-Whitney tests).[13] In each of the six panels at least 76% of the data are consistent with separation. In line with theory, conditional on type s, the distribution over actions is bimodal, placing substantial weight on action L, the optimal action conditional on identifying type s.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	**17%**	7%	**23%**
t	7%	9%	**37%**

First 10 Rounds

	L	C	R
s	**18%**	7%	**28%**
t	5%	10%	**32%**

Last 10 Rounds

(a) Mediated-Declaratives

	L	C	R
s	**20%**	6%	**25%**
t	5%	9%	**35%**

First 10 Rounds

	L	C	R
s	**20%**	9%	**24%**
t	4%	11%	**32%**

Last 10 Rounds

(b) Mediated-Directives

	L	C	R
s	**17%**	5%	**24%**
t	5%	7%	**42%**

First 10 Rounds

	L	C	R
s	**16%**	6%	**28%**
t	4%	7%	**39%**

Last 10 Rounds

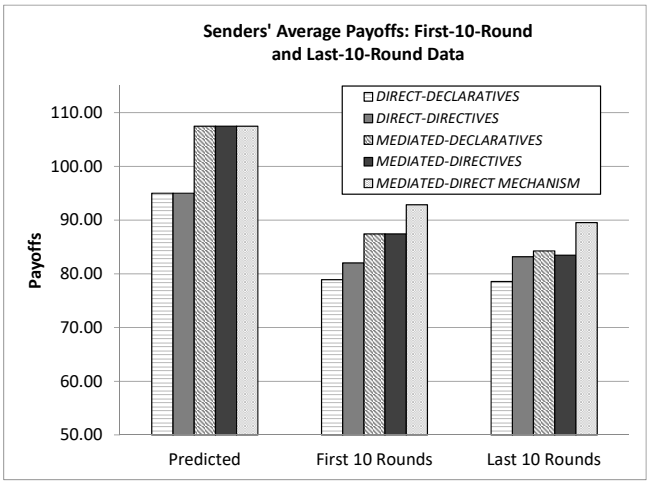
(c) Mediated-Direct Mechanism

Table 27: Mediated-Communication Outcomes (Treatment Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

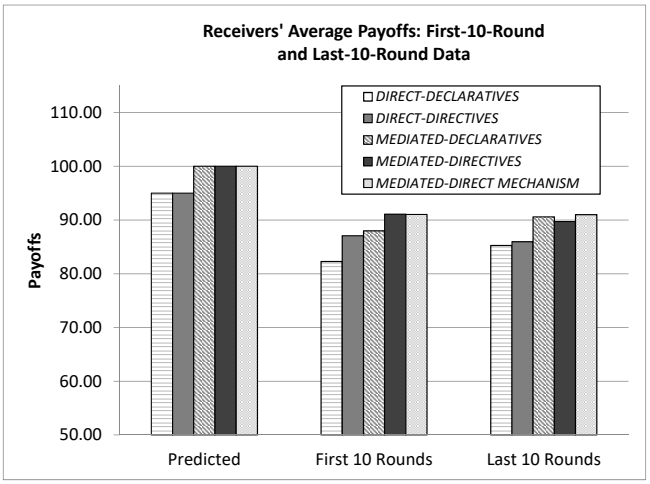
D.4 Individual treatments: payoffs

Figure 9 reports the payoffs for all five treatments over the first and last 10 rounds.

[13] For each of the four comparisons across treatments, a Kruskal-Wallis test further confirms that the three



(a) Senders' Average Payoffs



(b) Receivers' Average Payoffs

Figure 9: Average Payoffs in Direct Talk vs. Mediated Talk: First-10-Round and Last-10-Round Data

Payoff differences are small. This is not that surprising given that predicted payoff differences are themselves small. It is even less surprising in light of the fact that there is substantial noise in both sender and receiver behavior. Combining sender and receiver behavior compounds that noise when considering outcomes and payoffs. Nevertheless, regardless of whether we consider the first ten or the last ten rounds, and for both senders and receivers, payoffs in the mediated-talk treatments are never less than payoffs in the direct talk treatments. This suggests that mediation has a positive effect on payoffs.

frequencies have no statistical differences from one another ($p \geq 0.2535$).

E Appendix: Session-level data

In this section we report outcome data for each session, for the first and last 10 rounds.

E.1 Session level data on sender behavior

Figure 10 reports sender behavior in each of the five direct-declaratives sessions. There is some heterogeneity: in Session 2 types s send message "t" at three times the rate they send it in all other sessions. Also, in the first ten rounds there are three sessions in which types t send message "t" with considerably higher frequency than in the other two sessions; this behavior is consistent with over-communication in the first ten rounds of those three sessions. Overall, however, there is uniformity in modal behavior. In all five sessions, in both the first 10 and the last 10 rounds, and for both types the modal message is "s", consistent with the level-k prediction.

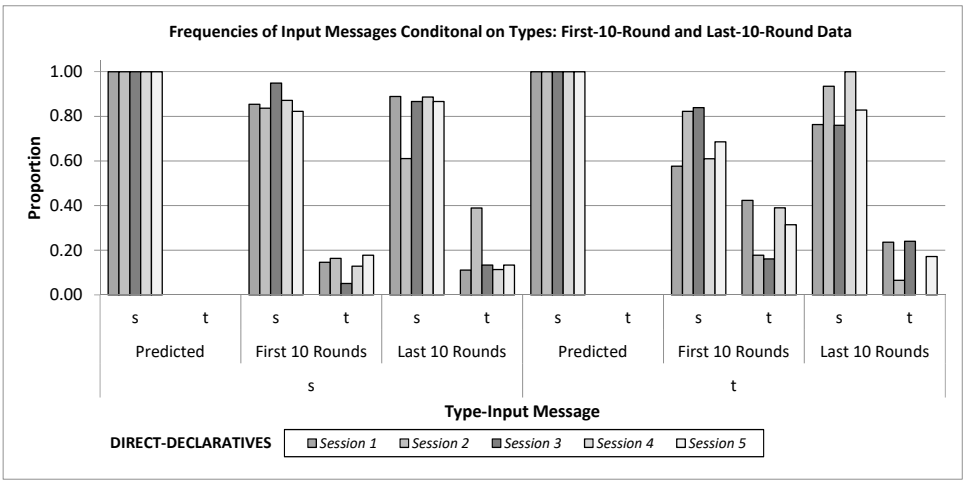


Figure 10: Senders' Behavior in Direct-Declaratives: First-10-Round and Last-10-Round Data

Figure 11 reports sender behavior in each of the five direct-directives sessions. There is considerable over-communication in the first ten rounds: in four of the five sessions more than 40% of types t send message "R"; while this is consistent with postulated level-0 behavior, predicted behavior for all higher levels is for types t to send message "L". This over-communication disappears in the last ten rounds. There modal behavior is uniform across sessions: both types in all sessions send message "L" with at least probability 0.8. This is in line with the level-k prediction.

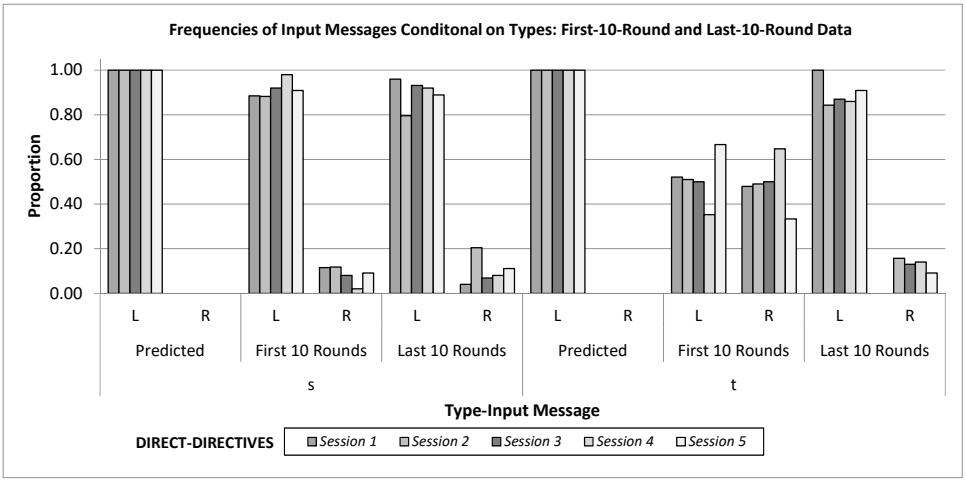


Figure 11: Senders' Behavior in Direct-Directives: First-10-Round and Last-10-Round Data

Figure 12 reports sender behavior in each of the five mediated-declaratives sessions. There is some under-communication in both the first and last ten rounds. Even in the last ten rounds there are two sessions in which type t senders send message "s" at least 30% of the time. Modal behavior is uniform across both the first and last ten periods and across all sessions: the majority of types s send message "s" and the majority of types t send message "t". This is the separating strategy predicted by the level-k analysis.

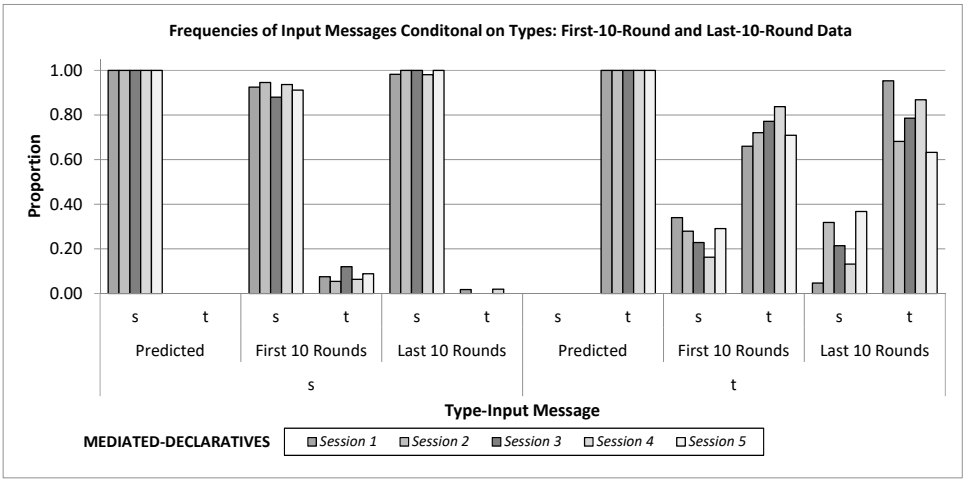


Figure 12: Senders' Behavior in Mediated-Declaratives: First-10-Round and Last-10-Round Data

Figure 13 reports sender behavior in each of the five mediated-directives sessions. There is some under-communication and heterogeneity, especially in the last ten rounds, with one session being fully separating and another session in which types t send the two messages "L" and "R" with roughly equal probability. In the first ten rounds modal behavior is separation in all five sessions. In the last ten rounds modal behavior is separation in four out of five sessions. Separation is predicted by the level-k analysis.

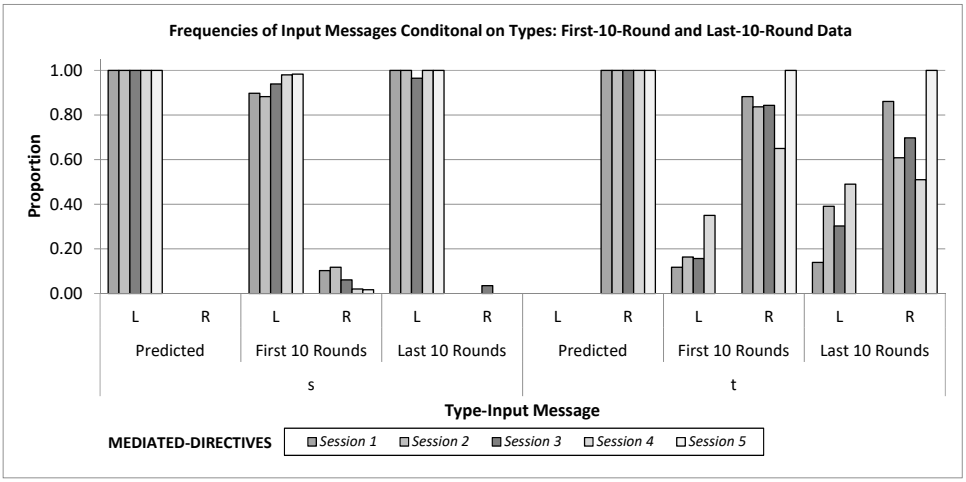


Figure 13: Senders' Behavior in Mediated-Directives: First-10-Round and Last-10-Round Data

Figure 14 reports sender behavior in each of the five mediated-direct-mechanism sessions. Again, there is some under-communication and heterogeneity, especially in the last ten rounds, with three sessions in which types t send message "L" more than 30% of the time. In the first ten rounds modal behavior is separation in all five sessions. In the last ten rounds modal behavior is separation in four out of five sessions. Separation is predicted by the level-k analysis.

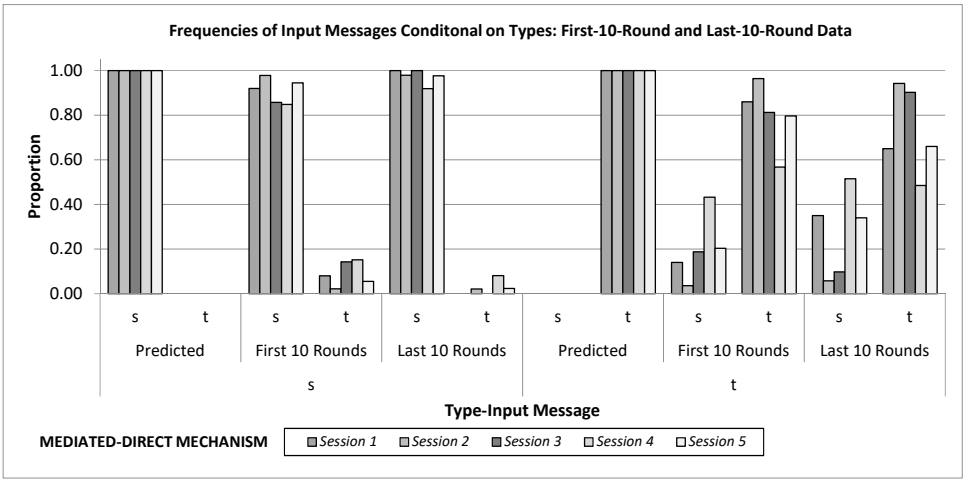


Figure 14: Senders' Behavior in Mediated-Direct Mechanism:
First-10-Round and Last-10-Round Data

E.2 Session level data on receiver behavior

Figure 15 reports receiver behavior in each of the five direct-declarative sessions. There is heterogeneity in the response to message "t" both in the first and in the last ten rounds. Message "t" should not be observed according to predicted sender behavior and is indeed observed relatively infrequently. In the first ten rounds in three of the sessions the modal response is C. In the last ten rounds action C is still a frequent response in two sessions, but the modal response is R in four out of five sessions. Both responses C and R to message "t" are consistent with the pooling equilibrium prediction in which only message "s" is sent, but only C is supported by the level-k analysis.

Message "s" is the only message receivers should observe according to predicted sender behavior and is also the most frequent message they do observe. In the first ten rounds the modal response to message "s" is action R, consistent with the level-k analysis, in four of the five sessions. The other frequent response to message "s" in the first ten round is action L, which is the best response to postulated level-0 sender behavior and would be a best response with sufficient over-communication by senders. The modal response to message "s" is action R in all five sessions in the last ten periods, consistent with predicted receiver behavior.

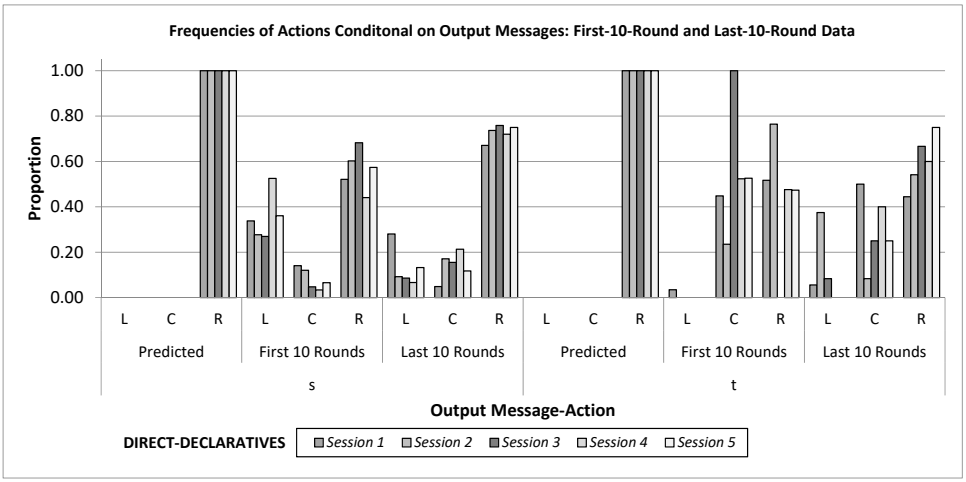


Figure 15: Receivers' Behavior in Direct-Declaratives: First-10-Round and Last-10-Round Data

Figure 16 reports receiver behavior in each of the five direct-directives sessions. There is heterogeneity in the response to message "R" both in the first and in the last ten rounds. Message "R" should not be observed according to predicted sender behavior and is indeed observed relatively infrequently by the receiver. In the first ten rounds in three of the sessions the receiver's modal response is C. In the last ten rounds action C is still a frequent response in two sessions, but the modal response is R in four out of five sessions. Both responses C and R to message "R" are consistent with the pooling equilibrium prediction in which only message "L" is sent, but only C is supported by the level-k analysis. This closely resembles the pattern we observe with responses to message "t" in the direct-declaratives sessions.

Message "L" is the only message receivers should observe according to predicted sender behavior and is also the most frequent message they do observe. In the first ten rounds the modal response to message "L" is action L in all five sessions. This is consistent with postulated level-0 behavior, but not with predicted level-k behavior for any level above 0; it would be a best reply with sufficient over-communication by senders. The modal response to message "s" is action R in all five sessions in the last ten periods, consistent with predicted receiver behavior. Thus in all five sessions there is a dramatic shift in behavior from the first ten rounds to the last ten rounds in the direction of the theoretical prediction.

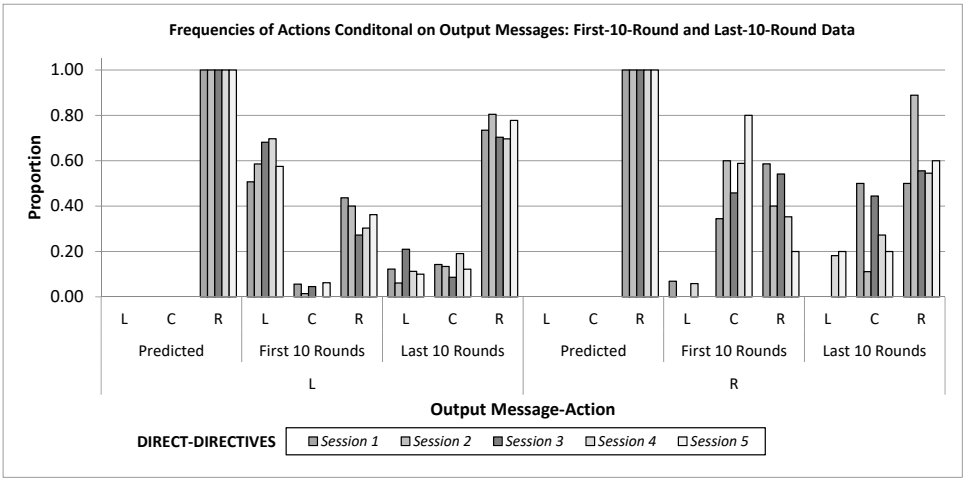


Figure 16: Receivers' Behavior in Direct-Directives: First-10-Round and Last-10-Round Data

Figure 17 reports receiver behavior in each of the five mediated-declaratives sessions. In the first ten rounds, in four out of five sessions, the modal receiver strategy is separation, responding to message "s" with action L and to message "t" with action R. In the last ten rounds there are two sessions in which the modal receiver strategy is separation; in two sessions the modal receiver strategy is pooling. Thus while we see separation more often than with direct talk, there is considerable heterogeneity and the tendency toward separation is more pronounced in the early rounds.

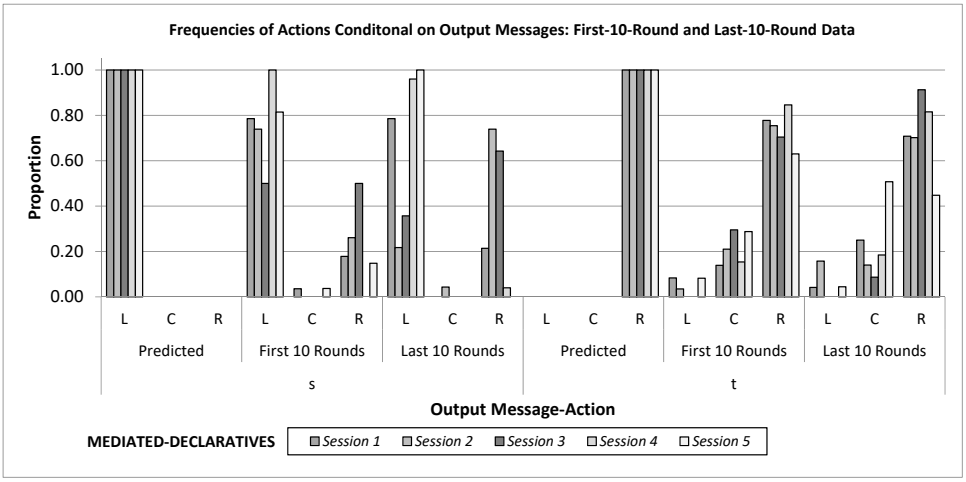


Figure 17: Receivers' Behavior in Mediated-Declaratives:
First-10-Round and Last-10-Round Data

Figure 18 reports receiver behavior in each of the five mediated-directives sessions. In the first ten rounds, in all five sessions, the modal receiver strategy is separation, responding to message "L" with action L and to message "R" with action R. In the last ten rounds there are four sessions in which the modal receiver strategy is separation. As with mediated declaratives, separation is more pronounced in the first ten than the last ten rounds. Departures from separation are in the direction of taking action R following message "L" and taking action C in response to message "R". The former suggests increased pessimism about being able to extract information form message "L", while the latter suggests, increased optimism about the ability to extract information from message "R". The tendency toward separation as the modal behavior is in line with predicted behavior.

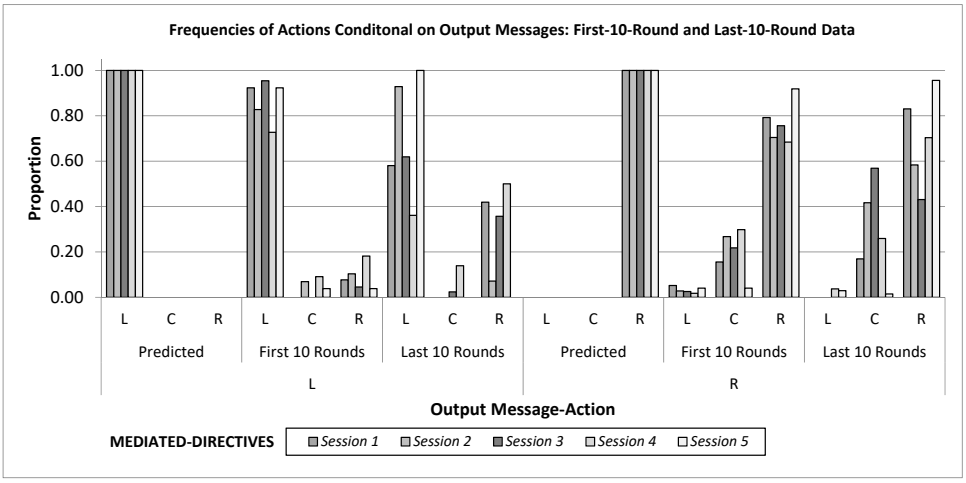


Figure 18: Receivers' Behavior in Mediated-Directives:
First-10-Round and Last-10-Round Data

Figure 19 reports receiver behavior in each of the five mediated-direct-mechanism sessions. In the first ten rounds, in all five sessions, the modal receiver strategy is separation, responding to message "L" with action L and to message "R" with action R. In the last ten rounds there are four sessions in which the modal receiver strategy is separation. As with mediated declaratives, separation is more pronounced in the first ten than the last ten rounds. Departures from separation are in the direction of taking action R following message "L" and taking action C in response to message "R". The former suggests increased pessimism about being able to extract information form message "L", while the latter suggests, increased optimism about the ability to extract information from message "R". The behavior pattern in the mediated-direct-mechanism sessions closely resembles that in the mediated directives sessions. The tendency toward separation as the modal behavior is in line with predicted behavior.

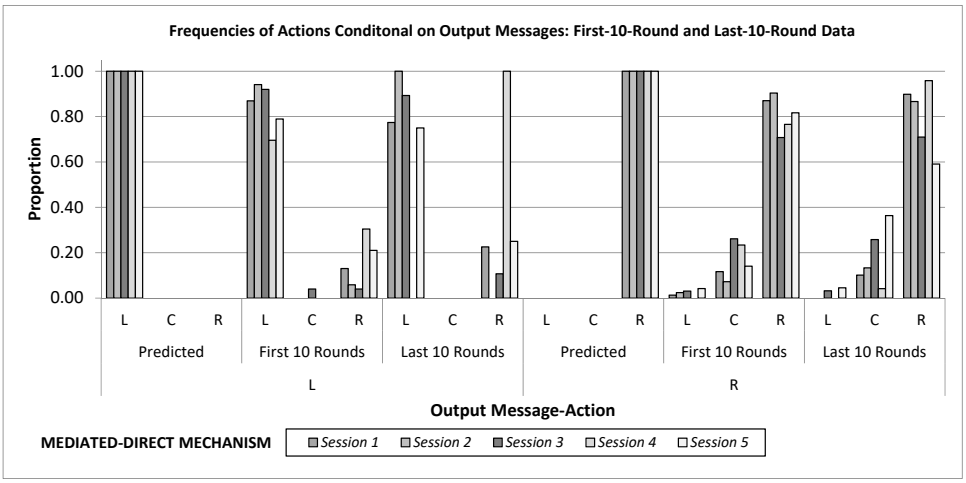


Figure 19: Receivers' Behavior in Mediated-Direct Mechanism: First-10-Round and Last-10-Round Data

E.3 Session-level outcomes: direct declaratives

Table 28 presents the observed outcomes for each of the five direct-declaratives sessions, aggregated over the first 10 and the last 10 rounds.

In all five sessions in the last ten rounds more than 60% of the data are consistent with pooling and in four out of five sessions more than 70% of the data are consistent with pooling. In each session there is more weight on the pooling outcome during the last ten rounds than during the first ten rounds. Session level data support the conclusion that in the direct-talk declaratives treatment behavior in the last ten rounds is best described by pooling.

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	10%	9%	**29%**
t	15%	14%	**23%**

First 10 Rounds

	L	C	R
s	10%	5%	**30%**
t	14%	8%	**33%**

Last 10 Rounds

(a) Session 1

	L	C	R
s	13%	7%	**35%**
t	10%	7%	**28%**

First 10 Rounds

	L	C	R
s	11%	7%	**36%**
t	5%	8%	**33%**

Last 10 Rounds

(b) Session 2

	L	C	R
s	13%	7%	**36%**
t	11%	7%	**26%**

First 10 Rounds

	L	C	R
s	4%	11%	**49%**
t	4%	6%	**26%**

Last 10 Rounds

(c) Session 3

	L	C	R
s	23%	5%	**21%**
t	16%	11%	**24%**

First 10 Rounds

	L	C	R
s	4%	12%	**38%**
t	3%	11%	**32%**

Last 10 Rounds

(d) Session 4

	L	C	R
s	16%	9%	**31%**
t	11%	9%	**24%**

First 10 Rounds

	L	C	R
s	6%	8%	**42%**
t	5%	6%	**33%**

Last 10 Rounds

(e) Session 5

Table 28: Communication Outcomes in Direct-Declaratives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

E.4 Session-level outcomes: direct directives

Table 29 presents the observed outcomes for each of the five direct-directives sessions, aggregated over the first 10 and the last 10 rounds.

In all five sessions in the last ten rounds more than 65% of the data are consistent with pooling, and the weight on pooling increases from the the first to the last ten rounds. During the first ten rounds there are systematic departures from pooling. In terms of our level-k analysis, the pattern of these departures from pooling is consistent with there being a mix of L_0 and of $L_{k\geq 1}$ players. According tho the level-k analysis, all type-action combinations except (s, C) have positive probability. This is consistent with (s, C) being the least frequently observed type-action pair in the first ten rounds of each of the five direct-talk directive sessions.

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	24%	4%	**24%**
t	14%	10%	**24%**

First 10 Rounds

	L	C	R
s	6%	6%	**38%**
t	6%	9%	**35%**

Last 10 Rounds

(a) Session 1

	L	C	R
s	28%	6%	**17%**
t	13%	13%	**23%**

First 10 Rounds

	L	C	R
s	4%	6%	**39%**
t	1%	7%	**43%**

Last 10 Rounds

(b) Session 2

	L	C	R
s	32%	3%	**21%**
t	18%	12%	**14%**

First 10 Rounds

	L	C	R
s	9%	8%	**32%**
t	10%	4%	**37%**

Last 10 Rounds

(c) Session 3

	L	C	R
s	34%	1%	**14%**
t	14%	19%	**18%**

First 10 Rounds

	L	C	R
s	8%	8%	**34%**
t	4%	12%	**34%**

Last 10 Rounds

(d) Session 4

	L	C	R
s	29%	8%	**18%**
t	17%	13%	**15%**

First 10 Rounds

	L	C	R
s	2%	7%	**36%**
t	9%	6%	**40%**

Last 10 Rounds

(e) Session 5

Table 29: Communication Outcomes in Direct-Directives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

E.5 Session-level outcomes: mediated declaratives

Table 30 presents the observed outcomes for each of the five mediated declaratives sessions, aggregated over the first 10 and the last 10 rounds.

If we measure proximity to separation versus pooling according to whether the frequency of (s, L) outcome realizations is closer to 25% (as separation would predict) than to 0% (as pooling would predict), then there are three sessions in which the outcome in the last ten periods is closer to separation than to pooling. During the first ten rounds four sessions would be categorized as separating according to this role of thumb. Finally, in all five sessions in the last ten rounds (s, L) is the most frequent type-action pair in which action R is not taken.

With only three out of five sessions closer to separation than pooling, there is considerable heterogeneity across sessions in the last ten rounds. Still, the modal outcome realization not involving action R is (s, L) in all five sessions in the last ten rounds and in four out of five sessions in the first ten rounds, suggesting an overall tendency toward separation.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	**17%**	8%	**28%**
t	11%	3%	**33%**

First 10 Rounds

	L	C	R
s	**23%**	10%	**24%**
t	2%	8%	**33%**

Last 10 Rounds

(a) Session 1

	L	C	R
s	**15%**	6%	**25%**
t	9%	9%	**36%**

First 10 Rounds

	L	C	R
s	**11%**	5%	**29%**
t	6%	6%	**43%**

Last 10 Rounds

(b) Session 2

	L	C	R
s	**8%**	10%	**23%**
t	5%	12%	**42%**

First 10 Rounds

	L	C	R
s	**7%**	3%	**43%**
t	2%	3%	**42%**

Last 10 Rounds

(c) Session 3

	L	C	R
s	**26%**	3%	**23%**
t	2%	8%	**38%**

First 10 Rounds

	L	C	R
s	**21%**	7%	**29%**
t	6%	7%	**30%**

Last 10 Rounds

(d) Session 4

	L	C	R
s	**21%**	7%	**17%**
t	7%	15%	**33%**

First 10 Rounds

	L	C	R
s	**27%**	9%	**15%**
t	9%	25%	**15%**

Last 10 Rounds

(e) Session 5

Table 30: Communication Outcomes in Mediated-Declaratives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

E.6 Session-level outcomes: mediated directives

Table 31 presents the observed outcomes for each of the five mediated directives sessions, aggregated over the first 10 and the last 10 rounds.

If we measure proximity to separation versus pooling according to whether the frequency of (s, L) outcome realizations is closer to 25% than to 0% then there are four sessions in which the outcome in the last ten periods is closer to separation than to pooling. The session closer to pooling is Session 4. There are also four out of five sessions closer to separation than pooling during the first ten rounds, and three sessions are closer to separation than pooling throughout.

With four out of five sessions closer to separation than pooling, there is heterogeneity across sessions in the last ten rounds. Still, the modal outcome realization not involving action R is (s, L) in four out of five sessions in the last ten rounds and in all five sessions in the first ten rounds, suggesting an overall tendency toward separation.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	**10%**	4%	**29%**
t	8%	9%	**40%**

First 10 Rounds

	L	C	R
s	**19%**	4%	**29%**
t	1%	7%	**40%**

Last 10 Rounds

(a) Session 1

	L	C	R
s	**23%**	9%	**19%**
t	3%	12%	**34%**

First 10 Rounds

	L	C	R
s	**20%**	14%	**20%**
t	6%	16%	**24%**

Last 10 Rounds

(b) Session 2

	L	C	R
s	**19%**	7%	**23%**
t	4%	10%	**37%**

First 10 Rounds

	L	C	R
s	**19%**	16%	**22%**
t	7%	18%	**18%**

Last 10 Rounds

(c) Session 3

	L	C	R
s	**21%**	9%	**26%**
t	7%	13%	**24%**

First 10 Rounds

	L	C	R
s	**10%**	10%	**26%**
t	7%	11%	**36%**

Last 10 Rounds

(d) Session 4

	L	C	R
s	**25%**	3%	**30%**
t	2%	1%	**39%**

First 10 Rounds

	L	C	R
s	**33%**	0%	**25%**
t	1%	1%	**40%**

Last 10 Rounds

(e) Session 5

Table 31: Communication Outcomes in Mediated-Directives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

E.7 Session-level outcomes: mediated direct mechanism

Table 32 presents the observed outcomes for each of the five mediated-talk direct mechanism sessions, aggregated over the first 10 and the last 10 rounds.

If we measure proximity to separation versus pooling according to whether the frequency of (s, L) outcome realizations is closer to 25% than to 0% then there are four sessions in which the outcome in the last ten rounds is closer to separation than to pooling. The session (very) near to pooling is Session 4. All five sessions are closer to separation than pooling during the first ten rounds.

With four out of five sessions closer to separation than pooling, there is heterogeneity across sessions in the last ten rounds. Still, the modal outcome realization not involving action R is (s, L) in four out of five sessions in the last ten rounds and in all five sessions in the first ten rounds, suggesting an overall tendency toward separation.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	**16%**	4%	**30%**
t	5%	5%	**40%**

First 10 Rounds

	L	C	R
s	**19%**	5%	**36%**
t	5%	2%	**33%**

Last 10 Rounds

(a) Session 1

	L	C	R
s	**14%**	4%	**27%**
t	4%	2%	**49%**

First 10 Rounds

	L	C	R
s	**22%**	4%	**22%**
t	3%	6%	**43%**

Last 10 Rounds

(b) Session 2

	L	C	R
s	**23%**	7%	**17%**
t	4%	13%	**36%**

First 10 Rounds

	L	C	R
s	**26%**	6%	**12%**
t	4%	12%	**40%**

Last 10 Rounds

(c) Session 3

	L	C	R
s	**17%**	6%	**24%**
t	6%	10%	**37%**

First 10 Rounds

	L	C	R
s	**0%**	0%	**53%**
t	0%	3%	**44%**

Last 10 Rounds

(d) Session 4

	L	C	R
s	**16%**	4%	**20%**
t	4%	7%	**49%**

First 10 Rounds

	L	C	R
s	**16%**	14%	**18%**
t	8%	12%	**32%**

Last 10 Rounds

(e) Session 5

Table 32: Communication Outcomes in Mediated-Direct Mechanism (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds